
# Recueil des résumés longs

*55èmes Journées de Statistique de la SFdS*

*à Bordeaux du 27 au 30 mai 2024*

Université de Bordeaux, Campus Victoire
3ter Pl. de la Victoire, 33000 Bordeaux

jds2024@u-bordeaux.fr

**Comité d'organisation :** Vincent Couallier et Robin Genuer ; **Comité scientifique :** Anne Ruiz-Gazen

# Table des matières

# Sessions invitées semi-plénières

# Statistical needs for Exposome Analytics: an Illustrative overview

Marc Chadeau-Hyam*[1]

[1]School of Public Health, Dept of Epidemiology and Biostatistics, Imperial College, London – Royaume-Uni

**Résumé**

The Exposome concept has been developed as a necessary complement to the genome to better understand the determinants of health and of the risk of chronic diseases. The external exposome combines a large range of external stressors (i.e. non-genetic) factors potentially impacting human health from conception onwards. These external exposures (i) are heterogeneous in nature, scale, and variability, (ii) feature complex correlation patterns and (iii) may operate as mixtures. The internal exposome can be defined as the way these exposures are embodied and its exploration relies on the screening and integration of high-resolution molecular data. While methods for omics data analyses are established, their application in an exposome context is raising specific methodological challenges including the analysis of complex and correlated exposures. Furthermore, the isolated exploration of an omic profile offers the possibility to capture stressor-induced biological/biochemical alterations, potentially impacting individual risk profiles, but this may only yield a fractional picture of the complex molecular events involved, therefore limiting our understanding of the effective mechanisms mediating the effect of the exposome. This defines three main methodological challenges in Exposome analytics: (i) reproducible and interpretable feature selection, (ii) data integration, and (iii) complexity reduction. Taking examples from real-world exposome projects we will illustrate the use of statistical and machine learning techniques to address these challenges and accommodate co-occurring exposures contributing to population stratification, explore the links between these and health outcomes, and investigate the (multi)-omic response to these sets of exposures.

**Mots-Clés:** Exposome, Omics Data, Stability selection, Graphical models

*Intervenant

# Statistique pour des trajectoires qualitatives : applications en analyse des données sensorielles

Hervé Cardot*[1]

[1]Université de Bourgogne – Université de Bourgogne-Franche-Comté – Dijon, France

**Résumé**

Ce travail est motivé par l'analyse de données sensorielles où on dispose de panels de trajectoires qualitatives issues d'expériences de dégustation. Deux questions importantes se posent.
Les dégustateurs, au niveau de la population, distinguent-ils deux produits (en terme de séquence temporelle des sensations lors de la dégustation de ces produits) ?

Peut on distinguer des "groupes homogènes" de dégustateurs ?

Un modèle intéressant et relativement simple pour ajuster les trajectoires qualitatives individuelles repose sur les processus semi-markoviens dont les paramètres peuvent être estimés par maximum de vraisemblance. Nous développons des techniques de type "two-sample test" pour répondre à la première question, via des permutations, tests de Wald et test du rapport de vraisemblance. Toujours en considérant la vraisemblance des trajectoires, une approche "model-based clustering", et une procédure d'estimation via un algorithme EM pénalisé, permet de segmenter les individus selon leur trajectoire.

Si le temps le permet, une approche concurrente basée sur une extension à temps continu de l'analyse factorielle des correspondances sera aussi présentée, avec ses avantages et ses défauts.

Travail réalisé en étroite collaboration avec Cindy Frascolla (IMB), Guillaume Lecuelle (Inrae CSGA), Caroline Peltier (Inrae CSGA), Pascal Schlich (Inrae CSGA) et Michel Visalli (Inrae CSGA), avec le soutien financier de la région BFC et l'Inrae.

**Mots-Clés:** processus semi markov, model based clustering, two sample test, algorithme EM, analyse sensorielle

*Intervenant

# Réseaux de gènes : inférence, évaluation, utilisation et au-delà

Nathalie Vialaneix [1]

[1] *Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet-Tolosan, France*

**Résumé.** La collecte de données à l'échelle moléculaire s'est considérablement accrue au cours des vingt dernières années, en quantité mais aussi en variété et précision. Cette évolution rapide crée un besoin de développement de méthodes d'analyse adaptées à la complexité et au volume de ces données : l'espoir pour les biologistes est que l'utilisation de ces nouvelles données ouvre la voie à une meilleure compréhension du fonctionnement du vivant et des relations complexes entre séquence d'ADN, environnement et ce que l'on peut observer à l'échelle de l'individu. Les répercussions potentielles sur le traitement des maladies (dont le cancer) ou la sélection des espèces agricoles animales et végétales pour faire face au changement climatique touchent à des questions sociétales importantes.

Une des données moléculaires les plus utilisées et étudiées pour caractériser le fonctionnement des cellules est l'*expression des gènes* (aussi appelée transcriptomique), qui est un mécanisme sous forte régulation génétique et épigénétique. Il est courant de représenter ces régulations sous la forme de graphes (ou réseaux) de gènes et la reconstruction de ces graphes, à partir de données d'expériences temporelles ou statiques, a été et demeure un sujet actif de recherche en statistique, connu sous le nom d'*inférence de réseaux de gènes* [11, 4, 7, 3, 10, 5, 6, 2, 8, 13]. À l'inverse, plusieurs méthodes de prédiction (régression ou classification) ont été développées pour inclure cette information de régulation sous forme de graphe et estimer à partir de celle-ci un phénotype mesuré à l'échelle de l'individu [12, 9].

Dans cet exposé, je dresserai un panorama des méthodes d'inférence de réseaux et de prédiction à base de graphes (en particulier des réseaux de neurones pour graphes [1]) et je discuterai les limites actuelles de leur utilisation ou de leur évaluation en regard de la complexité des mécanismes moléculaires modélisés.

*Cette présentation inclut des résultats de travaux publiés ou en cours, réalisés en collaboration avec Céline Brouard, Anne Goelzer, Raphaël Mourad et Vincent Rocher.*

**Mots-clés.** réseaux de gènes, transcriptomique, inférence de réseau, réseaux de neurones pour graphe

**Abstract.** Data collection at the molecular level has grown considerably during the last twenty years, not only in quantity but also in variety and and precision. This rapid evolution creates a need for the development of analysis methods adapted to the complexity and volume of these data. The hope for biologists is that the use of these new data will pave the way to a better understanding of how living organisms function and will unravel part of the complex relationships between DNA sequence, environment and what can be observed at the individual level. The potential repercussions include treatment of diseases (including

cancer) and the selection of agricultural plant and animal species able to cope with climate change, both being critical societal issues.

One of the most widely used and studied molecular data characterizing cell functioning is *gene expression* (also known as transcriptomics). Gene expression is a complex molecular mechanism that is a highly genetically and epigenetically regulated and it is a common practice to represent these regulations in the form of graphs (or networks) of genes. The reconstruction of these graphs using data from temporal or static experiments, has been and remains an active subject of statistical research, known as gene network inference [11, 4, 7, 3, 10, 5, 6, 2, 8, 13]. In addition, several prediction methods (regression or classification) have been developed to include these regulatory networks and to use them to better estimate a phenotype measured at the organism level [12, 9].

In this talk, I will give an overview of network inference and prediction methods based on graphs (and, in particular, on graph neural networks [1]). I will discuss the current limits of their use or evaluation with respect to the complexity of the molecular molecular mechanisms being modeled.

*This presentation includes published and ongoing works made in collaboration with Céline Brouard, Anne Goelzer, Raphaël Mourad, and Vincent Rocher.*

**Keywords.** gene networks, transcriptomics, network inference, graph neural networks

# Bibliographie

[1] Céline Brouard, Raphaël Mourad, and Nathalie Vialaneix. Should we really use graph neural networks for transcriptomic prediction? *Briefings in Bioinformatics*, 25(2):bbae027, 2024.

[2] Océane Cassan, Sophie Lèbre, and Antoine Martin. Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. *BMC Genomics*, 22:387, 2021.

[3] J. Chiquet, Y. Grandvalet, and C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[5] M. Gallopin, A. Rau, and F. Jaffrézic. A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS ONE*, 8(10), 2013.

[6] Johann S. Hawe, Fabian J. Theis, and Matthias Heinig. Inferring interaction networks from multi-omics data. *Frontiers in Genetics*, 10:535, 2019.

[7] Vân Anh. Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, 2010.

[8] Yoonjee Kang, Denis Thieffry, and Laura Cantini. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. *Frontiers in Genetics*, 12:362, 2021.

[9] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[10] Daniel Marbach, James C. Costello, Robert Küffner, Nicci Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, the DREAM5 Consortium, Manolis Kellis, and Gustavo Collins, James J.and Stolovitsky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012.

[11] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *Annals of Statistic*, 34(3):1436–1462, 2006.

[12] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35, 2007.

[13] Michael Saint-Antoine and Abhyudai Singh. Benchmarking gene regulatory network inference methods on simulated and experimental data. bioRxiv preprint, 2023.

# On the simulation of extreme events with neural networks

Stéphane Girard*†[1], Michael Allouche[2], and Emmanuel Gobet[3]

[1]Modeles statistiques bayesiens et des valeurs extremes pour donnees structurees et de grande dimension – Inria Grenoble - Rhône-Alpes, Laboratoire Jean Kuntzmann – France

[2]Kaiko [Paris] – non défini – France

[3]Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR7641 – France

**Résumé**

This work aims at investigating the use of generative methods based on neural networks to simulate extreme events. Although very popular, these methods are mainly invoked in empirical works. Therefore, providing theoretical guidelines for using such models in an extreme value context is of primal importance. To this end, we propose an overview of some generative methods dedicated to extremes, giving theoretical tips on their tail behaviour thanks to both extreme-value and copula tools. More specifically, we shall focus on a new parametrization for the generator of a Generative adversarial network (GAN) adapted to the heavy tail framework. An analysis of the uniform error between an extreme quantile and its GAN approximation is provided: We establish that the rate of convergence of the error is mainly driven by the second-order parameter of the data distribution. The above results are illustrated on simulated data and real financial data.

**Mots-Clés:** Extreme, value theory, neural networks, generative models

*Intervenant

†Auteur correspondant: stephane.girard@inria.fr

# Quelques réflexions statistiques sur l'apprentissage automatique inspiré de la physique

Claire Boyer*[1,2]

[1]Sorbonne Universite – LPSM – France
[2]Institut Universitaire de France – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – France

**Résumé**

L'apprentissage automatique informé par la physique combine l'expressivité des approches reposant sur les données avec l'interprétabilité des modèles physiques. Dans ce contexte, nous considérons un problème de régression général où le risque empirique est régularisé par une équation différentielle partielle, quantifiant ainsi une information d'incohérence physique. Nous prouvons que, pour les a priori différentiels linéaires, le problème peut être formulé comme une tâche de régression à noyau, ce qui donne un cadre rigoureux pour l'analyse du ML informé par la physique. En particulier, l'a priori physique peut contribuer à améliorer la convergence de l'estimateur. L'implémentation directe d'estimateurs à noyau informés par la physique peut être fastidieuse, et les praticiens ont souvent recours à des réseaux neuronaux informés par la physique (PINNs). Si le temps nous le permet, nous présenterons quelques recommandations statistiques pour une utilisation correcte des réseaux neuronaux informés par la physique.

**Mots-Clés:** Apprentissage informé par la physique, méthodes à noyaux, RKHS, réseaux de neurones

*Intervenant

# Modèles génératifs pour l'estimation de lois a posteriori. Applications aux problèmes inverses et aux méthodes SBI (Simulation-Based Inference)

Sylvain Le Corff*[1]

[1]Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Universite – France

**Résumé**

Les modèles génératifs basés sur le score (SGM), aussi connus sous le nom de modèles de diffusion, visent à estimer une distribution en estimant des fonctions de score à l'aide d'échantillons perturbés issus de la distribution cible. Ces méthodes ont permis l'obtention de résultats empiriques très impressionnants dans différents domaines complexes (traitement d'image, séries temporelles, etc.) et garantissant des performances au-delà des méthodes de l'état de l'art. Dans cet exposé, nous présenterons de nouvelles méthodes de simulation de lois a posteriori basées sur ces approches.
Tout d'abord, les modèles génératifs basés sur le score ont récemment été appliqués avec succès à différents problèmes inverses avec des applications par exemple en imagerie médicale. Dans ce cadre, nous pouvons exploiter la structure particulière de la loi a priori définie par le SGM pour définir une séquence de problèmes inverses intermédiaires. À mesure que le niveau de bruit diminue, les lois a posteriori de ces problèmes inverses se rapprochent de la loi cible du problème inverse initial. Pour échantillonner cette séquence de lois, nous proposons d'utiliser des méthodes de Monte Carlo séquentielles (SMC). L'algorithme proposé, MCGDiff, bénéficie de garanties théoriques pour la reconstruction des lois cibles et diverses simulations numériques illustrent qu'il est plus performant que les méthodes concurrentes lorsqu'il s'agit de traiter des problèmes inverses mal posés dans un cadre bayésien.

Par ailleurs nous nous intéresserons aux applications où des simulateurs complexes sont utilisés, ce qui signifie que la vraisemblance (loi des observations sachant les paramètres du modèle) de ces modèles est généralement difficile à calculer. L'inférence basée sur la simulation (SBI) se distingue dans ce contexte en ne nécessitant qu'un ensemble de données issues de simulations pour former des modèles génératifs capables d'approcher la distribution a posteriori des paramètres d'entrée du simulateur sachant une observation donnée. Nous considérerons un cadre "grande échelle" dans lequel de multiples observations sont disponibles et où l'on souhaite tirer parti de leurs informations partagées pour mieux déduire les paramètres du modèle. Nous présenterons une méthode s'appuyant sur les SGM permettant d'estimer la distribution a posteriori simplement en utilisant les informations du réseau de scores formé sur des observations individuelles. Nous comparerons notre méthode à des approches concurrentes récemment proposées et démontrerons sa supériorité en termes de stabilité numérique et de coût de calcul.

*Intervenant

**Mots-Clés:** Modèles génératifs, Modèles de diffusion, Score based models, Lois a posteriori, Simulation based inference

# Topological Data Analysis : extracting insights from the "shape" of data

Kathryn Hess Bellwald*[1]

[1]Brain and Mind Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (BMI EPFL) – Suisse

**Résumé**

The guiding philosophy of topological data analysis (TDA) is that the "shape" of a point cloud of data should reveal meaningful structure in the data. I will briefly sketch the theoretical foundations of TDA, describe some of its most frequently used and powerful methods, and conclude with examples of significant applications in cell biology, neuroscience, and material science.

**Mots-Clés:** topologie, neuroscience, biologie cellulaire, science des matériaux

*Intervenant

# Apprentissage statistique en sciences du climat : exemple des ondes internes de gravité.

Aurélie Fischer [1] & Sothea Has [2] & Riwal Plougonven [3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Université Paris Cité, aurelie.fischer@lpsm.paris*
[2] *Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, hassothea@lpsm.paris*
[3] *Laboratoire de Météorologie Dynamique, Ecole Polytechnique, riwal.plougonven@lmd.polytechnique.fr*

**Résumé.** Dans cet exposé, nous considérerons l'application de méthodes d'apprentissage statistique en sciences du climat : l'objectif est d'améliorer la connaissance et la description de processus physiques de petite échelle. Pour tenir compte des effets de ces processus de petite échelle, qui ne sont pas explicitement décrits dans les modèles de climat, il peut être utile d'intégrer, grâce à l'apprentissage statistique, les informations précises qui peuvent être fournies par des observations de ces processus. Les processus de petite échelle auxquels nous nous intéressons dans [1] sont les ondes internes de gravité, ondes dues au phénomène de gravité et à un contraste de densité sur la verticale. Les ondes de gravité jouent en effet un rôle crucial dans la circulation atmosphérique au-dessus de 15-20 km.

Les observations dont nous disposons sont des mesures par ballons superpressurisés stratosphériques, obtenues dans le cadre de la campagne Stratéole 2, projet franco-américain du Centre national d'études spatiales. Le comportement quasi-Lagrangien des ballons permet d'accéder à des estimations précises de flux de quantité de mouvement associés aux ondes de gravité dans la basse stratosphère. Les variables explicatives décrivant l'écoulement à grande échelle sont quant à elles fournies par les données de réanalyse ERA5 provenant du Centre Européen pour les Prévisions Météorologiques à Moyen Terme.

Ce travail a été effectué dans le cadre d'un projet de l'institut des Mathématiques pour la Planète Terre.

**Mots-clés.** Apprentissage statistique, Applications en sciences du climat.

**Abstract.** In this presentation, we will consider the application of statistical learning methods in climate science: the aim is to improve knowledge and description of small-scale physical processes. To take into account the effects of these small-scale processes, which are not explicitly described in climate models, it can be useful to integrate, through statistical learning, the precise information that may be provided by observations of these processes. The small-scale processes we focus on in [1] are internal gravity waves, caused by gravity and a density contrast on the vertical. Gravity waves indeed play a crucial role in atmospheric circulation above 15-20 km.

Our observations are based on measurements taken by stratospheric superpressure balloons, obtained as part of the Stratéole 2 campaign, a Franco-American project run by the Centre national d'études spatiales. The quasi-Lagrangian behavior of the balloons gives us access to precise estimates of momentum fluxes associated with gravity waves in the lower stratosphere. Explanatory variables describing the large-scale flow are provided by ERA5 reanalysis data from the European Centre for Medium-Range Weather Forecasts.

**Keywords.** Statistical learning, applications to climate sciences.

# Bibliographie

[1] S. Has, R. Plougonven, A. Fischer, R. Rani, F. Lott, A. Hertzog, A. Podglajen, M. Corcos (2024). Reconstructing balloon-observed gravity wave momentum fluxes using machine learning and input from ERA5, *Journal of Geophysical Research - Atmospheres*.

# Survival analysis of breast cancer screening programmes

Maja Pohar-Perme*[1] and Vratanar Bor[1]

[1]Institute of Biostatistics and Medical Informatics, Medical faculty, University of Ljubljana – Slovénie

**Résumé**

Cancer screening is a programme for medical screening of asymptomatic people who are at risk of developing cancer. In Slovenia, women between 50 and 70 are invited biannually to mammography screening. The programme has been running since 2008, we wish to evaluate its effectiveness. The most direct way to evaluate a screening programme is through survival analysis – we wish to know whether patients who participated in the programme have better chances of survival than those who did not take part. However, it turns out that any straightforward comparison of survival probabilities results in important biases that should not be neglected. In our work, we split the complex problem into simpler building blocks and show how survival can be compared in each of these blocks. While some of the issues can be solved non-parametrically, parametric assumptions may be needed for others. We have formulated a general theory and we adapt it to the particular issues of Slovene breast screening programme.

**Mots-Clés:** cancer screening programmes, lead time, length time bias, survival analysis

*Intervenant

# Curvature measures for random excursion sets: theoretical and computational developments

Elena Di Bernardino*[1]

[1]Laboratoire Jean Alexandre Dieudonné – LJAD – France

**Résumé**

The excursion set of a smooth random field carries relevant information in its various geometric measures. Geometric properties of these exceedance regions above a given level provide meaningful theoretical and statistical characterizations for random fields defined on Euclidean domains. Many theoretical results have been obtained for excursions of Gaussian processes and include expected values of the so-called Lipschitz{-Killing curvatures (LKCs), such as the area,
perimeter and Euler characteristic in two-dimensional Euclidean space. In this talk we will describe
a recent series of theoretical and computational contributions in this field.
Our aim is to provide answers to questions like:

- How the geometric measures of an excursion set can be inferred from a discrete sample of the
excursion set;

- How these measures can be related back to the distributional properties of the random field
from which the excursion set was obtained;

- How the excursion set geometry can be used to infer the extremal behavior of random fields

**Mots-Clés:** Statistique spatiale, champs aléatoires, extrêmes, statistique géométrique

*Intervenant

# Estimation non-paramétrique de l'intensité d'un processus ponctuel spatial par forêts aléatoires

Christophe Biscio[1] and Frédéric Lavancier*[2]

[1]Department of Mathematical Sciences [Aalborg] – Danemark
[2]Centre de Recherche en Economie et Statistique [Bruz] – Ensai, Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

**Résumé**

La fonction intensité d'un processus ponctuel spatial quantifie le nombre moyen de points par unité de mesure. Il s'agit typiquement de la première caractéristique d'intérêt que l'on cherche à estimer en présence de données réelles. Lorsqu'aucune co-variable n'est observée, son estimation se fait couramment par un estimateur à noyaux. Mais cette approche est mal adaptée à des domaines irréguliers (par exemple non-connexes), et sa consistance n'a lieu que si le nombre de points s'intensifie et pas si le domaine d'observation grandit. Lorsque des co-variables sont observées et que l'intensité en dépend, la situation est a priori plus favorable, mais les estimateurs à noyaux s'avèrent inefficaces : la plupart des études supposent alors une forme paramétrique log-linéaire de l'intensité, qui peut paraître contraignante. Dans ce travail, nous montrons comment adapter l'approche par forêts aléatoires à l'estimation non-paramétrique de l'intensité, avec ou sans présence de co-variables. Cela permet de gérer de façon satisfaisante des domaines non-réguliers et un grand nombre de co-variables, tout en mesurant leur influence. D'un point de vue théorique, lorsque les forêts sont purement aléatoires (ce qui est le cadre naturel sans co-variable), nous étudions la consistance de cette méthode lorsque l'on intensifie les points ou que l'on grandit le domaine d'observation, et ce pour une large classe de processus ponctuels. Nous montrons en particulier que la vitesse de convergence est en général plus rapide lorsqu'on s'appuie sur des co-variables et que l'intensité en dépend.

**Mots-Clés:** Statistique spatiale, Processus ponctuels spatiaux, Forêts aléatoires, Minimax

*Intervenant

# DPPs everywhere: repulsive point processes for Monte Carlo integration and machine learning

Rémi Bardenet*[1,2]

[1]CNRS – CNRS, CNRS : UMR8568, CNRS, CNRS : UMR6074, CNRS, CNRS : UMR5593, CNRS : ERL3189, CNRS : UMR7104, CNRS : UMR5244, CNRS : UMR2205, CNRS : UPR8241, CNRS, CNRS : UMR5554, CNRS : UMR5274, CNRS : UMR5493, CNRS : UMR7199, CNRS : UMR8184, CNRS : UMR7141, CNRS : UMR5251, CNRS : UMRTemps8066, CNRS : FR550, CNRS : UMR5127, CNRS : UMRSETEMoulis, CNRS : UMR9189 – France

[2]Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) - UMR 9189 – Ecole Centrale de Lille, Institut National de Recherche en Informatique et en Automatique, Institut Mines-Télécom [Paris], Université de Lille, Centre National de la Recherche Scientifique : UMR9189, CNRS : UMR9189 – France

**Résumé**

Determinantal point processes (DPPs) are specific repulsive point processes, which were introduced in the 1970s by Macchi to model fermion beams in quantum optics. More recently, they have been studied as models and sampling tools by statisticians and machine learners. Important statistical quantities associated to DPPs have geometric and algebraic interpretations, which makes them a fun object to study and a powerful algorithmic building block. After a quick introduction to determinantal point processes, I will discuss some of our recent statistical applications of DPPs in statistical sampling tasks.

**Mots-Clés:** processus ponctuels

*Intervenant

# Prix Marie-Jeanne Laurent Duhamel

# Apprentissage statistique de collections de réseaux avec applications en écologie et en sociologie

Saint-Clair Chabert-Liddell*[1]

[1]Mathématiques et Informatique Appliquées – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR0518 – France

**Résumé**

Les réseaux d'interactions constituent une façon naturelle de représenter sous forme de graphe les échanges ou relations existant entre un ensemble de nœuds représentant des espèces ou des individus. Considérer des collections de réseaux permet d'étudier des systèmes hétérogènes, composés de plusieurs sortes d'interactions impliquant différents types de nœuds. Lorsque les différents réseaux de la collection sont liés par une relation hiérarchique, nous parlerons de réseaux multiniveaux. Le modèle à blocs stochastiques a prouvé sa pertinence pour modéliser l'hétérogénéité du comportement des nœuds dans un unique réseau. Des extensions aux collections de réseaux et aux réseaux multiniveaux sont proposées. Elles permettent d'obtenir un clustering des nœuds des réseaux en fonction de leur rôle dans l'écosystème ou le système social, et de résumer la structure du système à l'échelle mésoscopique à travers un faible nombre de paramètres. L'inférence de ces modèles est complexe et des méthodes variationnelles sont adaptées à cette fin. Des méthodes de sélection de modèles permettent également de déterminer la dépendance entre les niveaux pour les réseaux multiniveaux et la similarité entre les structures pour les collections de réseaux. Les méthodes développées sont appliquées sur des données issues des sciences sociales et de l'écologie.

**Mots-Clés:** Graphe aléatoire, Réseaux d'interaction, Modèle à blocs stochastiques, Clustering

*Intervenant

# Modèles espace-état pour la prévision de séries temporelles. Application aux marchés électriques.

Joseph De Vilmarest*[1]

[1]Viking Conseil – Viking Conseil [Paris] – France

**Résumé**

Prévoir la demande électrique est fondamental pour la stabilité du réseau. En effet, l'électricité ne se stocke pas et l'équilibre entre consommation et production doit être maintenu en temps réel; en France, c'est la mission de RTE (Réseau de Transport d'Électricité). Ce travail a pour point de départ la conviction que notre monde est de plus en plus instable, tout particulièrement dans le secteur électrique. Citons comme exemples non exhaustifs : l'évolution des usages de l'électricité (e.g. l'introduction de véhicules électriques), l'augmentation des capacités de production non pilotable (énergies solaire et éolienne), et une instabilité croissante de la demande, en témoignent la crise du coronavirus en 2020 et la crise énergétique depuis 2022. Ces changements structurels motivent l'élaboration de modèles de prévision adaptatifs, en mesure de s'adapter aux changements de régime; c'est donc le principe de ce travail.
Nous nous sommes intéressés au cadre des modèles espace-état, et montrons qu'ils captent bien ces évolutions. Dans ce paradigme, nous représentons l'environnement par un état caché, dont dépend la demande à chaque instant. Nous estimons l'état par des méthodes bayésiennes telles le filtre de Kalman. Ces méthodes sont récursives; elles dépendent des observations à notre disposition et des hypothèses effectuées sur la dynamique des données. L'estimation de l'état nous permet d'obtenir une prévision de la demande.

Au cours de cette présentation, les modèles espace-état seront introduits dans le cadre de la prévision de séries temporelles. Puis nous verrons la calibration de ces modèles adaptatifs; elle consiste à choisir les variances du processus. Ces variances peuvent s'interpréter comme des vitesses d'évolution, et leur choix est similaire au choix du pas de gradient dans un algorithme de descente de gradient de second ordre. Enfin, l'application à la prévision de demande électrique sera illustrée sur les données de consommation nette (consommation réduite de productions non pilotables) en Grande-Bretagne.
Thèse réalisée à Sorbonne Université sous la direction d'Olivier Wintenberger, encadrée par Yannig Goude et Thi Thu Huong Hoang à EDF.

**Mots-Clés:** consommation électrique, modèle espace état, prévision de séries temporelles

*Intervenant

# Environnement et statistique

# Cooperative learning of Pl@ntNet's Artificial Intelligence algorithm using label aggregation

Tanguy Lefort [1] & Antoine Affouard [2] & Pierre Bonnet [3] &
Benjamin Charlier [4] & Alexis Joly [5] & Joseph Salmon [5]

[1] *Univ. Montpellier, CNRS, IMAG, Inria, LIRMM, France tanguy.lefort@umontpellier.fr*
[2] *IRD, AMAP, Montpellier, France antoine.affouard@cirad.fr*
[3] *CIRAD, AMAP, Montpellier, France*
[4] *Univ. Montpellier, CNRS, IMAG, France benjamin.charlier@umontpellier.fr*
[5] *Inria, LIRMM, France, alexis.joly@inria.fr*
[6] *Univ. Montpellier, CNRS, IMAG, IUF, France joseph.salmon@umontpellier.fr*

**Résumé.** Le système Pl@ntNet collecte des données à l'échelle mondiale en permettant aux utilisateurs de télécharger et d'annoter des observations de plantes. Les étiquettes ainsi obtenues bruitées en raison des compétences diverses des utilisateurs. L'obtention d'un consensus est cruciale pour entraîner des modèles d'apprentissage, mais l'échelle des données collectées rend les stratégies traditionnelles d'agrégation des étiquettes difficiles à mettre en œuvre. En outre, comme de nombreuses espèces sont rarement observées, l'expertise des utilisateurs ne peut pas être évaluée comme un accord entre utilisateurs : sinon, les experts en botanique auraient un poids plus faible dans l'étape d'apprentissage que l'utilisateur moyen de part leur participation moindre mais plus ciblée. La stratégie d'agrégation d'étiquettes que nous proposons vise à entraîner de manière coopérative des modèles d'apprentissage automatique pour l'identification des plantes. Cette stratégie estime l'expertise des utilisateurs sous la forme d'un score de confiance par travailleur, basé sur leur capacité à identifier des espèces végétales à partir de données collectées par la foule. Le score de confiance est estimé récursivement à partir des espèces correctement identifiées compte tenu des étiquettes estimées actuelles. Ce score interprétable exploite les connaissances des experts en botanique et l'hétérogénéité des utilisateurs. Nous évaluons notre stratégie sur un large sous-ensemble de la base de données Pl@ntNet axée sur la flore européenne, comprenant plus de 6 000 000 d'observations et 800 000 utilisateurs. Nous démontrons que l'estimation des compétences des utilisateurs basée sur la diversité de leur expertise améliore la performance de l'étiquetage.

**Mots-clés.** Apprentissage coopératif, aggrégation d'étiquettes, annotation de données, écologie

**Abstract.** The Pl@ntNet system enables global data collection by allowing users to upload and annotate plant observations, leading to noisy labels due to diverse user skills. Achieving consensus is crucial for training, but the vast scale of collected data makes traditional label aggregation strategies challenging. Additionally, as many species are rarely observed, user expertise can not be evaluated as an inter-user agreement: otherwise, botanical experts would have a lower weight in the training step than the average user as they have fewer but precise participation. Our proposed label aggregation strategy aims to cooperatively train plant identification models. This strategy estimates user expertise as a trust

score per worker based on their ability to identify plant species from crowdsourced data. The trust score is recursively estimated from correctly identified species given the current estimated labels. This interpretable score exploits botanical experts' knowledge and the heterogeneity of users. We evaluate our strategy on a large subset of the Pl@ntNet database focused on European flora, comprising over 6 000 000 observations and 800 000 users. We demonstrate that estimating users' skills based on the diversity of their expertise enhances labeling performance.

**Keywords.** Crowdsourcing, label aggregation, data annotation, ecology

# 1 Introduction

Computer vision models are a great aid in plant species recognition in the field [20, 1]. However, to train them we need large annotated datasets. These datasets are often created thanks to citizen science approaches, collecting both reliable and useful information [2]. Among existing plant recognition applications, the Pl@ntNet system enables global data collection by allowing users to upload and annotate plant observations.

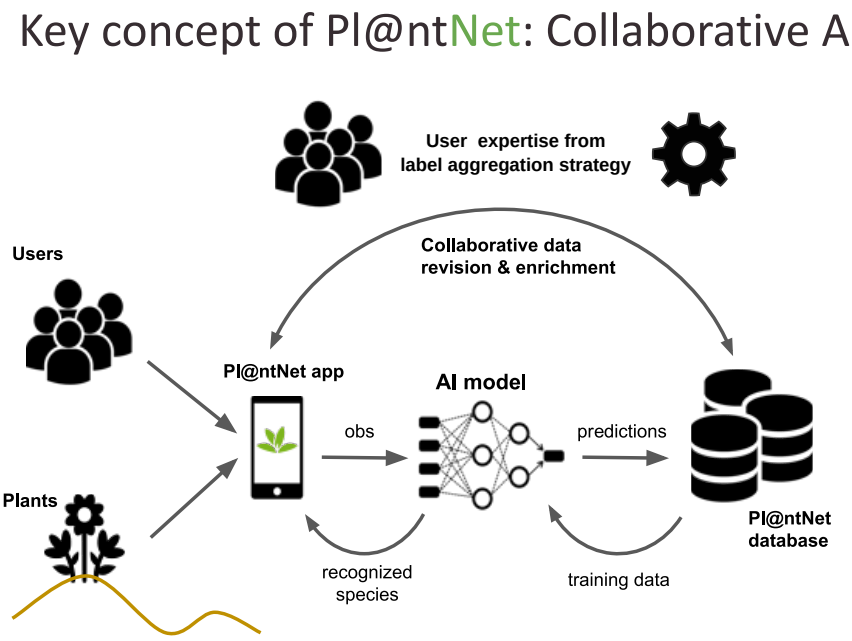



Figure 1: Pl@ntNet system for plant species recognition. Users take their plant observations in the Pl@ntNet application. A prediction is output by the neural network model. Users can validate the prediction or propose another species. The whole votes collection is used to evaluate user expertise (see Algorithm 1) and actively revise observations identifications.

At the time of writing, this participatory approach has resulted in the collection of over 20 million observations (image or group of images of a same plant), belonging to almost

46 000 species, by more than 6 million observers worldwide. The collaborative process of Pl@ntNet is described in Figure 1. The model interacts with the human decision by proposing possible species given an observation. For each returned species, using a similarity search, the Pl@ntNet system also shows similar pictures from the database. This lets users to visually check that their observation is likely to belong to a predicted species given the most similar observations. Such a visual control can be of help when flowering is not yet complete to compare two plants at similar growth stages. Plant species identification is a task that requires skills to recognize morphological traits (shapes, measurements, environments and specific characteristics). A large number of users with diverse skills have participated in gathering plant observations and helped improve the training dataset of our computer vision model. Their participation is based on votes that they can cast on others observations, or by the initial species determination of their observation. The quality of each vote is then processed by the algorithm presented in Section 2.2.

Other citizen science projects such as iNaturalist [19] or eBird [18] use a similar approach to collect data. However, each platform has its own label aggregation strategy. The iNaturalist project, with more than 2.5 million users, records the votes at different taxonomic levels. The resulting label is the aggregation of at least two votes on a species-level identification (or coarser or finer taxonomic level) and the taxon needs at least two-thirds of identifier agreements – in particular, all users have the same weight in the decision-making. Over time, a taxon can be further refined by the community or revoked. eBird handles taxon quality control by using a checklist in each region for observers. Quality verifications on the checklist are performed and, combined with user knowledge – the number of species and checklist submitted, number of flagged observations, further discussions with local experts – the observation taxon is accepted. The eBird project also showed that monitoring species accumulation from observers can help to sort their skills [10]. While they consider the species accumulation by hours spent on each collected observation, we propose a strategy that takes into account the entire history of observations of the observer.

In this article, we present the Pl@ntNet label aggregation strategy. Using a large-scale dataset of more than 6 million observations and 800 thousand users, we show that our strategy can improve the quality of the collected data, without removing every observation that was only labeled by single users. This work is ongoing and the dataset will be released with codes.

# 2 Methods

## 2.1 Dataset and notation

To compare the different label aggregation strategies on large-scale datasets, we consider a subset of the Pl@ntNet database focused on Southwestern European flora observations – Baleares, Corsica, France, Portugal, Sardegna and Spain – from 2017 to October 2023. In total, 9 005 108 votes are cast by $n_{\text{user}} = 823\,251$ users on 6 699 593 observations after cleaning steps. Those cleaning steps include filtering out identification votes with proposed plant species not available in the World Checklist of Vascular Plants (WCVP) [6]. Thanks

to Kew's Royal Botanical Garden, we adopted the Plants of the World Online [15] system with the *k-southwestern-europe*. Within this taxonomic checklist, we removed synonyms. However, there are plant species listed in *k-southwestern-europe* POWO that are not in the WCVP checklist. As there is a possible taxon ambiguity in this case – multiple species possible for a given synonym depending on the referential – we leave the proposed label untouched.

**Notation** In the following, denote $K = 11\,425$ the number of species within the dataset. We index the observations by $i \in [n_\bullet] = \{1, \dots, n_\bullet\}$ where $\mathcal{D}_\bullet$ is the considered dataset composed of $n_\bullet$ observations and their associated votes. For example, the full south-western european flora dataset from Pl@ntNet of $6\,699\,593$ observations is denoted $\mathcal{D}_{\text{SWE}}$. Other subsets are presented in Section 2.3. We write $\mathcal{U}$ the set of users. Each user $u$ has a unique identifier used as an index, and we denote $\mathcal{U}_i$ the set of users that have voted on observation $i$ – *i.e.* $\mathcal{U} = \cup_{i \in [n_{\text{SWE}}]} \mathcal{U}_i$. The vote of user $u$ on observation $i$ is denoted $y_i^u \in [K]$. Each observation $i$ is created by an author $u$ stored in Author($i$).

## 2.2 Proposed label aggregation strategy

Pl@ntNet label aggregation strategy relies on estimating the number of correctly identified species for each user. Similar to other strategies, we rely on an EM based iterative procedure [5] to estimate consecutively the users' skills and each observation's species. As the collected data is used to train the model, the label aggregation strategy also generates a trust indicator on the observation. This quality indicator reveals if the observation is valid or not. The AI model is then only trained on valid observations. This operation is done monthly to keep the system up-to-date with the latest observations. The more users vote on observations, the more valid observations are identified and the better the model. Notice that proposing a species as author of the observation weighs ten times more than voting by click in Algorithm 1. Indeed, being on the field leads to more information on the environment and a better determination of the species. Finally, note that species are unequivocally identified as author's ($n_u^{\text{author}}$ in Algorithm 1) or as votes on other's observations ($n_u^{\text{vote}}$) in the aggregation strategy. The final number of species identified by users is the aggregation of these two terms: $n_u = \text{Round}\left(n_u^{\text{author}} + \frac{1}{10} n_u^{\text{vote}}\right)$.

From Algorithm 1, we see that a user becomes **self-validating** (*i.e.* trusted enough so that their label checks observations as valid identifications) when their weight $w_u$ is greater than the level $\theta_{\text{conf}}$. In practice, this means that an experienced user who has collected enough weight can validate any observation without any other user's vote. Note that this identification can later be invalidated by other users with enough weight thanks to the accuracy threshold $\theta_{\text{conf}}$. Moreover, the weight function $f$ shown in Figure 2 is a non-decreasing function that maps the number of identified species $n_u$ to a trust score in the form of:

$$w_u = f(n_u) = n_u^\alpha - n_u^\beta + \gamma \ , \tag{1}$$

where $\alpha, \beta \in \mathbb{R}_+^\star$ are hyperparameters that were calibrated internally to fit prior knowledge and $\gamma > 0$ is the constant representing the initial weight of each user. In practice, we use

**Algorithm 1** Pl@ntNet label aggregation strategy

**Input:** Votes as $(u, y_i^u)_{i\in[n_{\text{SWE}}],u\in[n_{\text{user}}]}$ for each observation $i$ and user $u$ answering the voted species $y_i^u$, accuracy threshold $\theta_{\text{acc}}$, confidence threshold $\theta_{\text{conf}}$, weight function $f$, initial weight $\gamma > 0$

**Output:** Estimated labels $\hat{y}_i$ for each observation $i$

Initialize $\hat{y}_i = \text{MV}\left(\{y_i^u\}_u\right)$ for each observation $i \in [n_{\text{SWE}}]$

Initialize user weights as $w_u = \gamma$ for each user $u \in [n_{\text{user}}]$

**while** not converged **do**

  **for** each observation $i \in [n_{\text{SWE}}]$ **do**

    Compute label confidence: $\text{conf}_i(\hat{y}_i) = \sum_{u\in\mathcal{U}_i} w_u \mathbb{1}(y_i^u = \hat{y}_i)$

    Compute label accuracy: $\text{acc}_i(\hat{y}_i) = \text{conf}_i(\hat{y}_i) / \sum_{k\in[K]} \text{conf}_i(k)$

    Compute validity indicator: $s_i = \mathbb{1}(\text{acc}_i(\hat{y}_i) \geq \theta_{\text{acc}} \text{ and } \text{conf}_i(\hat{y}_i) \geq \theta_{\text{conf}})$

  **end for**

  **for** each user $u \in [n_{\text{user}}]$ **do**

    Compute the number of valid identified species for authoring observations:

$$n_u^{\text{author}} = |\{y_i^u \in [K] \,|\, y_i^u = \hat{y}_i, s_i = 1, \text{Author}(i) = u\}|$$

    Compute the number of identified species by voting on other's observations:

$$n_u^{\text{vote}} = |\{y_i^u \in [K] \,|\, y_i^u = \hat{y}_i, \text{Author}(i) \neq u\}|$$

    Compute the rounding number of identified species per user:

$$n_u = \text{Round}\left(n_u^{\text{author}} + \frac{1}{10} n_u^{\text{vote}}\right)$$

    Transform number of estimated species per user into trust score: $w_u = f(n_u)$

  **end for**

  Update estimated labels with a weighted majority vote

$$\forall i \in [n_{\text{SWE}}],\ \hat{y}_i = \underset{k\in[K]}{\arg\max} \sum_{u\in\mathcal{U}_i} w_u \mathbb{1}(y_i^u = k)$$

**end while**

$\alpha = 0.5$, $\beta = 0.2$ and $\gamma = \log(2.1) \simeq 0.74$ in the weight function. As for the two thresholds that control the level of uncertainty accepted for a given label, they are set to $\theta_{\text{conf}} = 2$ to control the total weight on an observation and $\theta_{\text{acc}} = 0.7$ to control the agreement between users given their expertise.

## 2.3 Evaluation against other aggregation strategies

**Existing aggregation strategies** Plant species label aggregation is a challenging task due to the large number of species $K$. Hence, many classical strategies in the label aggregation

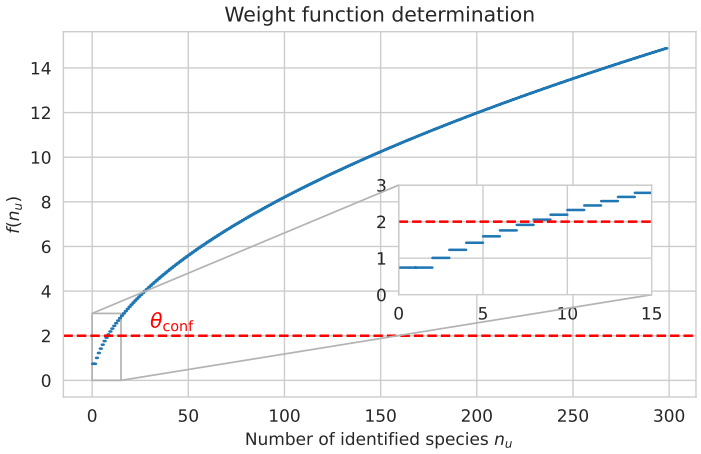



Figure 2: Weight function (Equation (1)) used to map the number of identified species to a trust score in the Pl@ntNet label aggregation strategy. The user confidence threshold $\theta_{\text{conf}} = 2$ requires a user to have identified at least $n_u = 8$ species to become **self-validating**. A new user starts with a weight of $f(0) = f(1) = \gamma \simeq 0.74$.

literature such as Dawid and Skene's [4] and other variations [14, 17] are not applicable as they require estimating a $K^2$ matrix for each worker. This would result, in the considered dataset $\mathcal{D}_{\text{SWE}}$ as $11\,425^2 \times 823\,251 \approx 10^{14}$ parameters to estimate. Similar issues occur for other label aggregation strategies [21, 8, 13]. We do not consider deep-learning based crowdsourcing strategies as Rodrigues and Pereira [16] and Chu, Ma, and Wang [3] or Lefort et al. [11] as they train a model from crowdsourced labels but do not output aggregated labels on the training set. In the Pl@ntNet application, we need to propose one or multiple species for each observation to users. To overcome these issues, we consider the following label aggregation strategies that can scale to large $K$ and number of users:

- **Majority Vote (MV)[9]**: Certainly the most common aggregation strategy, the majority vote selects the most answered label. In the case of equalities, a random draw is performed – creating sometimes some variability in the labeling process. More formally, given an observation $i$:

$$\text{MV}(i, \{y_i^u\}_u) = \arg\max_{k \in [K]} \sum_{u \in \mathcal{U}_i} \mathbb{1}(y_i^u = k) \quad .$$

- **Worker agreement with aggregate (WAWA)** [12]: Also known as the inter-rater agreement, this strategy weights each user by how much they agree with the MV labels on the images they annotated. More formally, given an observation $i$:

$$\text{WAWA}(i, \mathcal{D}_{\text{SWE}}) = \arg\max_{k \in [K]} \sum_{u \in \mathcal{U}_i} w_u \mathbb{1}(y_i^u = k)$$

$$\text{with } w_u = \frac{1}{|\{y_{i'}^u\}_{i'}|} \sum_{i'=1}^{n_{\text{SWE}}} \mathbb{1}\left(y_{i'}^u = \text{MV}(\{y_{i'}^u\}_u)\right) \quad .$$

- **iNaturalist** [19]: The iNaturalist platform generates a label for observations with at least two votes. The estimated label represents the one with at least two-thirds of the majority in agreement. Every user has the same weight in the aggregation. More formally:

$$\text{iNaturalist}(i, \{y_i^u\}_u) = \begin{cases} \text{MV}(i, \{y_i^u\}_u) & \text{if } s_i = 1 \\ \text{undefined} & \text{otherwise} \end{cases}$$
$$\text{with } s_i = \mathbb{1}\left( \max_{k \in [K]} \frac{1}{|\mathcal{U}_i|} \sum_{u \in \mathcal{U}_i} \mathbb{1}(y_i^u = k) \geq \frac{2}{3} \right) \ .$$

As there is no observation filter for the MV and WAWA, we consider that for all observation $i$, $s_i = 1$ for these two strategies. Experiments were completed using the `peerannot` library (`https://github.com/peerannot/peerannot`).

**Creation of an evaluation set in a crowdsourcing setting** To evaluate the performance of a label aggregation strategy, it is necessary to know the ground truth. However, in the context of crowdsourced data, there is no known truth for the observations. The sheer volume of data makes it impossible to ask botanical experts to create such ground truth for the whole database.

Instead of asking such experts to label a subset of the data, we identified botanical experts in our users database. From within the Pl@ntNet team, we referenced well-known botanists to start a list of expert users. To these we have added TelaBotanica [7] users with registered confirmed botanical experience from their directory and that are also Pl@ntNet users that participated to the South-Western Europe flora subset. Among the users, 98 are identified as botanical experts by the Pl@ntNet team and Telabotanica platform. The answers of these experts are considered as ground truth labels and used to evaluate strategies performance. Despite our selection process of supposedly "indisputable" experts, a few observations in the test set denoted $\mathcal{D}_{\text{expert}}$ still end up with contradictory labels (4 observations in total). As they represent a very small percentage, we simply removed them from $\mathcal{D}_{\text{expert}}$.

Our evaluation set $\mathcal{D}_{\text{expert}}$ is finally composed of 26 811 observations. Of these evaluation data, 17 125 received more than two identifications and are stored in $\mathcal{D}_{\text{multiple votes}}$; 1 263 have more than two votes with at least one disagreement between users are stored in $\mathcal{D}_{\text{disagreement}}$. Figure 3 shows the distribution of observations from $\mathcal{D}_{\text{SWE}}$ to the finer and more ambiguous $\mathcal{D}_{\text{disagreement}}$.

Unfortunately, the demand for multiple labels on observations is not being met, despite the large number of users. Indeed, 310 564 users were single time voters (meaning they interacted with the system only once).

**Evaluation metric** To evaluate the label aggregation strategies, we use the following accuracy metrics computed on valid observations ($s_i = 1$):

$$\text{Acc}(\hat{y}, y; \mathcal{D}_\bullet) = \frac{1}{n_\bullet} \sum_{i=1}^{n_\bullet} \mathbb{1}(\hat{y}_i = y_i) \mathbb{1}(s_i = 1) \ ,$$

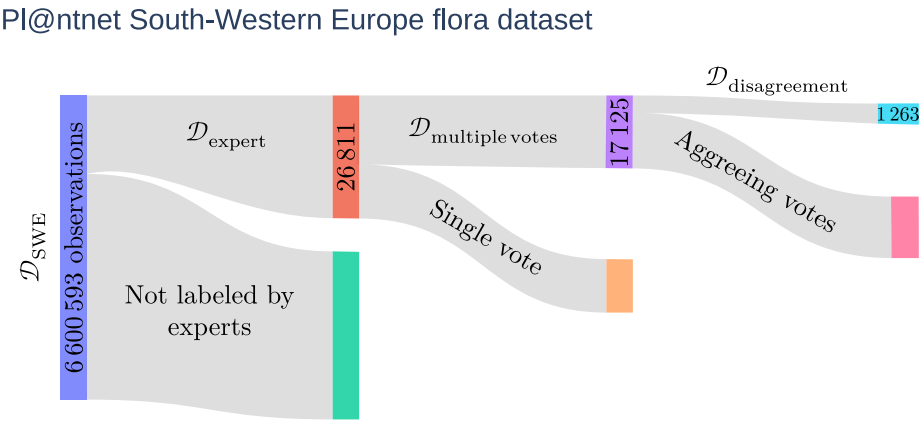


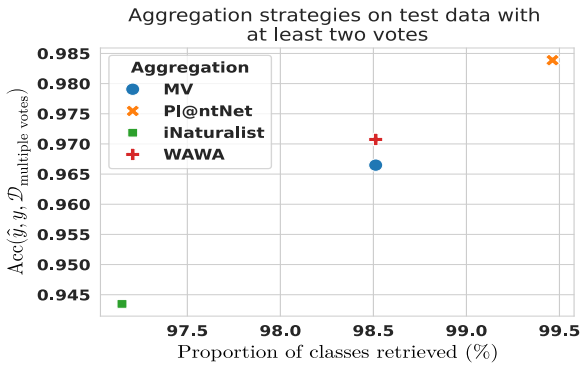

Figure 3: Log-scales distribution of the observations in the South-West European Flora subset from the Pl@ntNet database.

with $\hat{y} = (\hat{y}_i)_i$ the estimated labels on the considered set $\mathcal{D}_\bullet \subset \mathcal{D}_{\text{expert}}$, $y = (y_i)_i$ the associated experts labels, considered as ground truth. When the aggregation strategy indicates the observation as invalid ($s_i = 0$ for Pl@ntNet and iNaturalist), we consider the label as incorrect in the performance measure as an expert was able to decide on a species. Finally, we also consider the proportion of species retrieved by the aggregation strategies. This is important as if a species identified by the experts disappears during the aggregation, the model trained from this aggregated data can no longer predict this species.

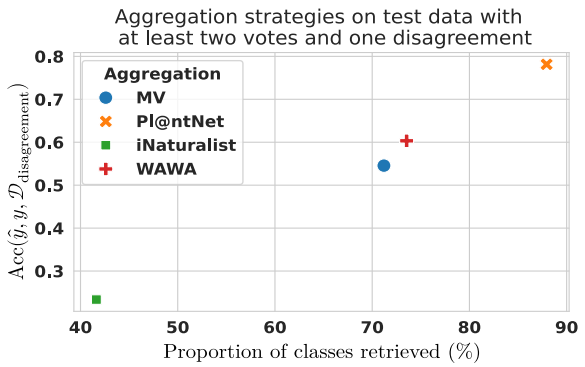
We evaluate the label recovery of each strategy on three subsets visualized in Figure 3: the full test set where experts have voted a species, the subset of observations with at least 2 votes and the subset of observations with at least 2 votes and one disagreement. The latter subset is the most challenging as it contains the observations with the most ambiguity. We selected these subsets to investigate the label aggregation strategies' performance depending on the ambiguity level.

# 3 Results

**Accuracy of the aggregation strategies** We evaluate the accuracy of the strategies on the set of tasks labeled by our experts. Figure 4 shows how many predicted labels match our experts answers on $\mathcal{D}_{\text{multiple votes}}$ and $\mathcal{D}_{\text{disagreement}}$. More importantly, we compare this quantity with the volume of class retrieved by the aggregation strategy. We observe that the data filtering from the iNaturalist strategy impacts its performance. On $\mathcal{D}_{\text{expert}}$, MV reaches 97% of accuracy, WAWA 98%, iNaturalist 60% and Pl@ntNet 99%. To differentiate between the best performing strategies, we need to look at more ambiguous observations like those in $\mathcal{D}_{\text{multiple votes}}$ and $\mathcal{D}_{\text{disagreement}}$. In high ambiguous frameworks, the WAWA strategy outperforms the MV one. However, overall the Pl@ntNet aggregation is more often in adequation with the experts and retrieves almost 90% of plant species identified by experts in high ambiguous datasets against 73% for WAWA, 71% for MV and only 41% for iNaturalist.




(a) Accuracy on $\mathcal{D}_{\text{multiple votes}}$ against volume of species recovered




(b) Accuracy on $\mathcal{D}_{\text{disagreement}}$ against volume of species recovered

Figure 4: Accuracy of the aggregation strategies against the volume of class retrieved on subsets with at least two votes – either agreeing (A) or with at least one disagreeing vote (B). The Pl@ntNet aggregation is more accurate especially in a highly ambiguous setting (B). The iNaturalist data filter highly impacts how many classes are kept in the dataset and the overall accuracy in both settings. WAWA and MV perform similarly with a benefit for WAWA when skill evaluation is needed.

# 4 Conclusion

We demonstrated that collaborative identification of plant species can effectively be used to obtain expert levels labels. Using a large subset of millions of observations and thousands of users from the Pl@ntNet organization, we investigate a label aggregation strategy that weighs user answers based on their estimated number of species correctly identified without using prior expert knowledge. Many strategies used previously either do not scale to the magnitude of the current databases – either Pl@ntNet, iNaturalist or eBird – or are outperformed by our aggregation. Our strategy weighs users based on the number of correctly identified species. This weight is interpretable and shows the diversity of the user's skillset. It can be directly applied on other crowdsourced frameworks with a high number of classes like iNaturalist's.

# References

[1] M. L. Borowiec et al. "Deep learning as a tool for ecology and evolution". In: *Methods in Ecology and Evolution* 13.8 (2022), pp. 1640–1660.

[2] E. D. Brown and B. K. Williams. "The potential for citizen science to produce reliable and useful information in ecology". In: *Conservation Biology* 33.3 (2019), pp. 561–569.

[3] Z. Chu, J. Ma, and H. Wang. "Learning from Crowds by Modeling Common Confusions." In: *AAAI*. 2021, pp. 5832–5840.

[4] A. P. Dawid and A. M. Skene. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *J. R. Stat. Soc. Ser. C. Appl. Stat.* 28.1 (1979), pp. 20–28.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 39.1 (1977), pp. 1–22.

[6] R. Govaerts. *The World Checklist of Vascular Plants (WCVP)*. Checklist dataset. Accessed via GBIF.org on 2024-01-30. 2023. DOI: 10.15468/6h8ucr.

[7] L. Heaton, F. Millerand, and S. Proulx. “« Tela Botanica » : une fertilisation croisée des amateurs et des experts”. In: *Hermès, La Revue* 57.2 (2010), pp. 61–68.

[8] D. Hovy et al. “Learning Whom to Trust with MACE”. In: *Proceedings of NAACL-HLT 2013*. 2013.

[9] G. James. “Majority vote classifiers: theory and applications”. PhD thesis. Stanford University, 1998.

[10] S. Kelling et al. “Can Observation Skills of Citizen Scientists Be Estimated Using Species Accumulation Curves?” In: *PLOS ONE* 10.10 (Oct. 2015), pp. 1–20.

[11] T. Lefort et al. “Identify ambiguous tasks combining crowdsourced labels by weighting Areas Under the Margin”. In: *arXiv preprint arXiv:2209.15380* (2022).

[12] A. Limited. *Calculating Worker Agreement with Aggregate (Wawa)*. 2021. URL: https://success.appen.com/hc/en-us/articles/202703205-Calculating-Worker-Agreement-with-Aggregate-Wawa-.

[13] Q. Ma and A. Olshevsky. “Adversarial crowdsourcing through robust rank-one matrix completion”. In: *NeurIPS*. Vol. 33. 2020, pp. 21841–21852.

[14] R. J. Passonneau and B. Carpenter. “The Benefits of a Model of Annotation”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 311–326.

[15] POWO. *Plants of the World Online*. Published on the Internet. Facilitated by the Royal Botanic Gardens, Kew. 2024. URL: http://www.plantsoftheworldonline.org/.

[16] F. Rodrigues and F. Pereira. “Deep learning from crowds”. In: *AAAI*. Vol. 32. 2018.

[17] V. B. Sinha, S. Rao, and V. N. Balasubramanian. “Fast Dawid-Skene: A fast vote aggregation scheme for sentiment classification”. In: *arXiv preprint arXiv:1803.02781* (2018).

[18] B. Sullivan et al. “eBird: A citizen-based bird observation network in the biological sciences”. In: *Biological conservation* 142.10 (2009), pp. 2282–2292.

[19] G. Van Horn et al. “The inaturalist species classification and detection dataset”. In: *CVPR*. 2018, pp. 8769–8778.

[20] M. Vidal et al. “Perspectives on individual animal identification from biology and computer vision”. In: *Integrative and comparative biology* 61.3 (2021), pp. 900–916.

[21] J. Whitehill et al. “Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise”. In: *NeurIPS*. Vol. 22. 2009. (Visited on 2021).

# Températures maximales en France au 21ème siècle

Occitane Barbaux [1], Philippe Naveau [2], Nathalie Bertrand[3]& Aurélien Ribes[4]

[1] *Institut de Radioprotection et de Sûreté Nucléaire, Centre National de Recherches Météorologiques, France, occitane.barbaux@umr-cnrm.fr*
[2] *Laboratoire des Sciences du Climat et de l'Environnement, Centre national de la recherche scientifique, France, philippe.naveau@lsce.ipsl.fr*
[3] *Institut de Radioprotection et de Sûreté Nucléaire, France, nathalie.bertrand@irsn.fr*
[4] *Centre National de Recherches Météorologiques, Université de Toulouse, Météo France, Centre national de la recherche scientifique, France, aurelien.ribes@meteo.fr*

**Résumé.** : Extrêmes de Température en France au 21ème siècle

Dans un contexte de changement climatique, il est nécessaire d'adapter les normes de protections contre divers aléas naturels. Ce travail vise à quantifier le risque de dépassement de niveaux élevés de température, à l'échelle d'un site d'intérêt et à l'horizon 2100.
Le niveau de retour est l'indicateur de niveau de risque usuel. Étant défini par sa probabilité de dépassement annuel, il n'est pas adapté pour l'étude du risque sur une période en l'absence de stationnarité. Un autre indicateur, capable de caractériser le risque sur l'ensemble d'une période d'intérêt, est nécessaire. La Fiabilité Équivalente, qui caractérise le maximum sur une période, a donc été sélectionnée car sa probabilité de dépassement sur une période est identique à la probabilité totale de dépassement sur la même période pour le niveau de retour (Liang 2016 et Hu 2018).
La Fiabilité équivalente est estimée en adaptant la méthode statistique de Ribes (2020) et Robin (2020).
La méthode des Maxima Annuels a été utilisée avec une distribution d'extremum généralisée (GEV) non stationnaire, appliquée à des maxima quotidien de température. La non-stationnarité sur les paramètres est ici donnée par une covariable, la température moyenne annuelle européenne, qui agit comme un proxy du réchauffement effectif.
L'utilisation d'un cadre bayésien nous permet d'intégrer diverses sources d'informations. Ainsi l'incertitude sur l'évolution future du climat et des divers scénarios est intégrée en formant une distribution a-priori à partir des trajectoires des Modèles de Climat Globaux. L'information sur les conditions locales du point d'intérêt est fournie lors de la contrainte par les données de mesures météorologiques sur site.
Après avoir comparé divers algorithmes de Monte-Carlo par chaînes de Markov, nous avons utilisé l'algorithme No-U-Turn Sampler (Holman 2014) implémenté par STAN pour estimer les distributions a-posteriori de nos paramètres.
L'estimateur prédictif intègre l'incertitude sur l'ensemble de la distribution pour chacun des paramètres. Il permet de prendre en compte une gamme plus large de valeurs possibles que l'estimateur médian. De plus, il permet d'obtenir une valeur unique, plus adaptée aux normes de sûreté que la notion d'intervalle de confiance. Nous l'avons ici adapté à la Fiabilité Équivalente.

Une application de notre méthode est faite pour la Vallée du Rhône, en France, sur la période 2050-2100 pour divers scénarios d'émissions.

**Mots-clés.** Environnement et statistique, Extrêmes et risques, Statistique bayésienne

**Abstract.** Extreme Temperature in France by 2100
In a context of climate change, design levels for environmental extremes have to be redefined. This work aims to estimate the risk of extreme temperature levels excess by 2100 at a local scale.
The usual risk indicator, the annual return level, is only defined in a stationary context. Since it's defined as a level corresponding to an annual probability of excess, its value can't be defined for a period of interest. It is necessary to select another risk indicator able to cover a full time period. The Equivalent Reliability, which defines the distribution of the maximum event during the period, was chosen for because its probability of excess over a period is the same as the total probability of excess over a period for a stationary return level (Liang 2016 et Hu 2018).
The method used to estimate the Equivalent Reliability is adapted from the statistical method by Ribes (2020) and Robin (2020).
An Annual Maxima framework and a non-stationary Generalized extreme value distribution, applied to daily maximum temperatures, are used to estimate extreme temperatures. The non stationarity in the parameters is given by a covariate, the European annual mean temperature, as a proxy of climate warming.
The method's Bayesian framework uses a prior probability distribution based on Global Climate Models to account for uncertainty on the future of climate change, which is then constrained by local meteorological measurements. After comparing several Markov chain Monte Carlo algorithm (MCMC), an estimation of the posterior parameters' distributions was produced using the No-U-Turn Sampler MCMC algorithm implemented in Stan (Holman 2014).
The predictive estimate, including uncertainty over the parameters' full distribution, was adapted to Equivalent Reliability. It accounts for a larger range of possible extreme values than the median estimate, while providing a unique value fit for design calculation.
Our method was applied to the Rhone Valley on France over the period 2050-2100 given several emissions scenarios.

**Keywords.** Environment and statistics, Extremes and risk, Bayesian statistics

# 1 Introduction

Le changement climatique entraîne une augmentation de l'intensité et de la fréquence des évènements de température extrêmes. Il est donc nécessaire d'adapter les normes de protections contre cet aléa naturel. Ce travail vise à quantifier le risque de dépassement de niveaux élevés de température, à l'échelle d'un site d'intérêt et à l'horizon 2100.
Pour cela, il est tout d'abord nécessaire de définir un indicateur de niveau de risque adapté

au contexte non-stationnaire: La fiabilité Équivalente. Cet indicateur est ensuite estimé à partir d'une méthode bayésienne qui permet l'intégration de deux sources d'informations: Les modèles Climatiques Globaux et les observations météorologiques locales. Enfin, l'incertitude sur les paramètres est intégrée dans une adaptation prédictive de l'indicateur.

# 2 La Fiabilité équivalente

Le niveau de retour est l'indicateur de niveau de risque usuel, notament pour les normes de construction. Étant défini par sa probabilité de dépassement annuel, il n'est pas adapté pour l'étude du risque sur une période en l'absence de stationnarité, puisque la probabilité annuelle d'un niveau de température fixe évolue chaque année. Il n'est donc plus possible de l'utiliser pour fixer une norme unique.
Plusieurs indicateurs tels que l'Expected Waiting Time ou l'Average Design Life Level ont été développés pour adapter la notion de niveau de retour au contexte non-stationnaire.
Notre problème nécessite un indicateur capable de caractériser le risque sur l'ensemble d'une période d'intérêt, tout en conservant une logique commune avec le niveau de retour, basé sur la probabilité annuelle de dépassement. La Fiabilité Équivalente, qui caractérise le maximum sur une période, a été sélectionnée car sa probabilité de dépassement sur une période est identique à la probabilité totale de dépassement sur la même période pour le niveau de retour (Liang 2016 et Hu 2018).
On l'obtient en résolvant l'équation suivante :

$$P[Max_{t\in[\mathbf{T_1},\mathbf{T_2}]}(Y_t) \leq \mathbf{z}^{\mathbf{ER}}_{\mathbf{T_2}-\mathbf{T_1}}] = (1-\frac{1}{\mathbf{T}})^{\mathbf{T_2}-\mathbf{T_1}+1}$$

Avec $T_1 - T_2$ la période d'intérêt, $Y_t$ le maximum annuel pour l'année $t$, et $T$ la période de retour équivalente.
Il est donc nécessaire d'estimer la distribution des maximuma sur une période.

# 3 Méthode d'estimation

Notre travail vise à estimer des températures rares et extrêmes. Nous avons donc utilisé la méthode des Maxima Annuels avec une distribution d'extremum généralisée (GEV) non stationnaire. Le paramètre de forme est le seul paramètre conservé stationnaire.

La non-stationnarité sur les paramètres est donnée par une covariable, la température moyenne annuelle européenne, qui agit comme un proxy du réchauffement effectif. Contrairement à une dépendance linéaire directe au temps, elle permet d'intégrer les possibles évolutions suivant les différents scénarios.

## 3.1 Intégration de l'information globale

Les distributions des paramètres de la GEV sont estimées en adaptant la méthode statistique de Ribes (2020) et Robin (2020).
L'utilisation d'un cadre bayésien nous permet d'intégrer diverses sources d'informations. Ainsi, l'incertitude sur l'évolution future du climat et des divers scénarios est intégrée en formant une distribution gaussienne multivariée a-priori à partir des trajectoires des Modèles de Climat Globaux. Dans notre application, nous disposons de 28 modèles de Climat Globaux disposant de 250 ans de données, de 1850 à 2100. Cette plage de temps corresponds à la période historique puis à la période future pour les Modèles de Climat Globaux.

## 3.2 Estimation des paramètres contraints

L'information sur les conditions locales du point d'intérêt est fournie lors de l'étape de contrainte par les données de mesures météorologiques d'une station Météo France locale. Dans notre application, les données de maxima quotidiens sont fournies par la station Météorologique de Pierrelatte. Elle est située à proximité du point d'intérêt, et sa chronique dure depuis 37 ans avec peu de ruptures.
Après avoir comparé divers algorithmes de Monte-Carlo par chaînes de Markov, nous avons choisi d'utiliser l'algorithme No-U-Turn Sampler (Holman 2014) implémenté par STAN pour estimer les distributions a-posteriori de nos paramètres.

# 4 Estimateur prédictif

L'estimateur prédictif intègre l'incertitude sur l'ensemble de la distribution pour chacun des paramètres. Il permet de prendre en compte une gamme plus large de valeurs possibles que l'estimateur médian. De plus, il permet d'obtenir une valeur unique, plus adaptée aux normes de sûreté que la notion d'intervalle de confiance.
Nous l'avons ici adapté à la Fiabilité Équivalente. L'indicateur est obtenu en résolvant :

$$P[Max_{t\in[\mathbf{T_1},\mathbf{T_2}]}(Y_t) \leq \mathbf{z}^{\mathbf{ER}}_{\mathbf{T_2}-\mathbf{T_1}} | Y^0, X^0] = (1 - \frac{1}{\mathbf{T}})^{\mathbf{T_2}-\mathbf{T_1}+1}$$

Avec $Y^0, X^0$ les observations locales et de la covariable.

La valeur visée est obtenue en tirant un grand nombre de maxima sur la période pour un grand nombre de tirages de jeux de paramètres GEV, puis en prenant le quantile de la distribution mélange ainsi obtenue.
Les distributions obtenues sont illustrées dans la figure ci-dessous:

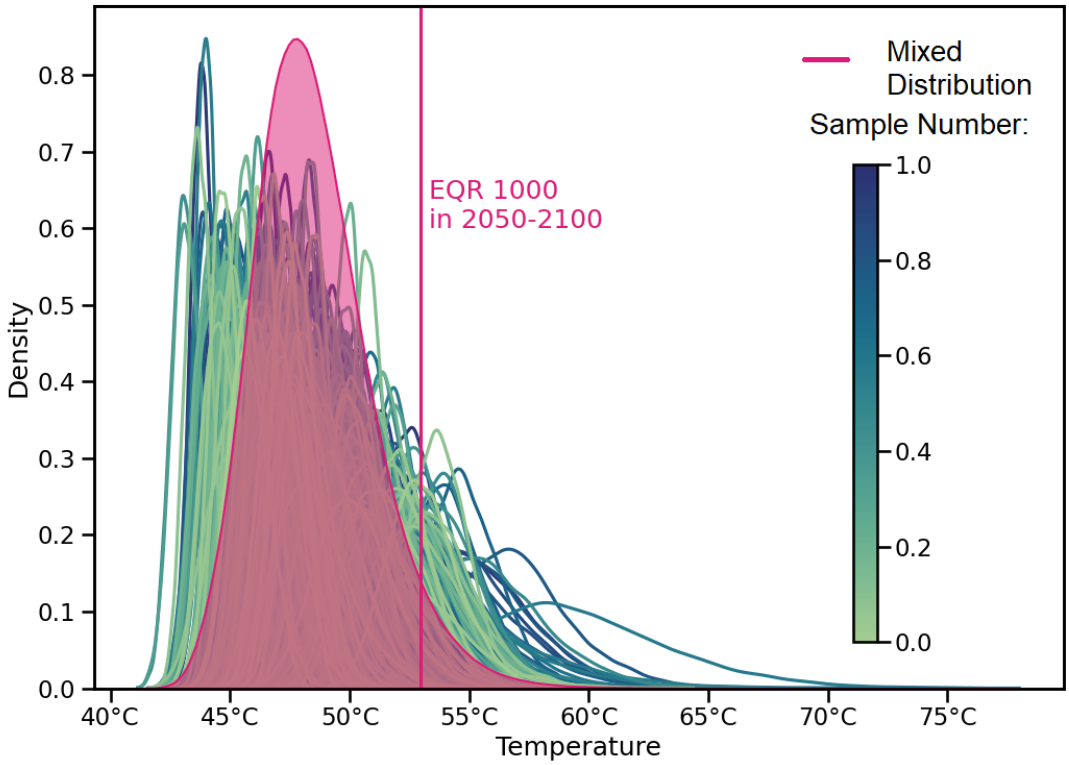



Figure 1: Distribution des maxima sur la période 2050-2100. Les distributions de bleu à vert correspondent à la distribution pour chaque unique jeu de paramètres. La distribution rouge est la distribution mélange pour l'ensemble des jeux de paramètres tirés.

Ainsi, dans l'exemple ci-dessus, avec une projection future de type scénario SSP 5-8.5 , on peut estimer que sur la période 2050-2100, la température qui a une probabilité annuelle de 0.001 et totale de 0.05 d'être atteinte ou dépassée est 53°C.

# Bibliographie

Homan, M.D. and Gelman, A. (2014) 'The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo', *The Journal of Machine Learning Research*, 15(1), pp. 1593-1623.

Hu, Y. et al. (2018) 'Concept of Equivalent Reliability for Estimating the Design Flood under Non-stationary Conditions', *Water Resources Management*, 32(3), pp. 997-1011.

Liang, Z. et al. (2016) 'Study on the estimation of design value under non-stationary environment', *South-to-North Water Transfers Water Sci Tech*, 14, pp. 50-53.

Ribes, A., Thao, S. and Cattiaux, J. (2020) 'Describing the Relationship between a Weather Event and Climate Change: A New Statistical Approach', *Journal of Climate*, 33(15), pp. 6297-6314.

Robin, Y. and Ribes, A. (2020) 'Nonstationary extreme value analysis for event attribution combining climate models and observations', *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(2), pp. 205-221.

# Analyse de l'impact de variables environnementales sur les réseaux plantes-pollinisateurs à l'aide d'auto-encodeurs variationnels pour graphes bipartites.

Emre Anakok[1] & Pierre Barbillon [2] & Colin Fontaine [3] & Elisa Thebault [4]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, emre.anakok@agroparistech.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, pierre.barbillon@agroparistech.fr*
[3] *Centre d'Écologie et des Sciences de la Conservation, MNHN, CNRS, SU, 43 rue Buffon, 75005 Paris, France, colin.fontaine@mnhn.fr*
[4] *Sorbonne Université, CNRS, IRD, INRAE, Université Paris Est Créteil, Université Paris Cité, Institute of Ecology and Environmental Sciences (iEES-Paris), 75005 Paris, France, elisa.thebault@upmc.fr*

**Résumé.** Nous proposons une méthode de représentation des réseaux bipartites à l'aide d'auto-encodeurs de graphes adaptés à l'étude des réseaux écologiques issus de données de sciences participatives. Ceci représente un double défis, puisque l'on veut mettre en évidence les effets de nombreuses covariables d'intérêt écologique, comme par exemple la dégradation des habitats, tout en prenant en compte les effets d'échantillonnage, notamment l'effet observateur. Nous adaptons l'approche de l'auto-encodeur variationnel de graphes au cas bipartite pour générer des représentations dans un espace latent où les deux ensembles de nœuds sont positionnés en fonction de leur probabilité de connexion. En incorporant le critère d'indépendance de Hilbert-Schmidt (HSIC) comme un terme de pénalité supplémentaire dans la perte que nous optimisons, nous nous assurons que la structure de l'espace latent est indépendante des variables continues, qui sont liées au processus d'échantillonnage. Nous appliquons notre méthode à l'ensemble de données Spipoll, un programme d'observation participatif des interactions entre plantes et pollinisateurs à travers toute la France auquel contribuent de nombreux observateurs, ce qui le rend biaisé car les participants sont sujet à un phénomène d'apprentissage au fur et a mesure de leur participation. Enfin, nous prédisons les changement de structure du réseau de pollinisation en fonction de variations de composition du paysage, avec ou sans prise en compte de l'experience des observateurs. Les resultats mettent en lumière l'importance de la correction des biais d'echantillonnage, avec par exemple, une connectivité du réseau largement augmenté dans les paysages agricole dominés par de l'elevage lorsque les biais d'échantillonnage sont corrigés.

**Mots-clés.** Graphes et réseaux, Réseaux de neurones, Échantillonnage, Statistique appliquée à l'écologie

**Abstract.** We propose a method to represent bipartite networks using graph embeddings tailored to tackle the challenges of studying ecological networks derived from citizen science data. This represents a twofold challenge, since we want to highlight the effects of numerous

covariates of ecological interest, such as habitat degradation, while taking into account sampling effects, in particular the observer effect. We adapt the variational graph auto-encoder approach to the bipartite case, which enables us to generate embeddings in a latent space where the two sets of nodes are positioned based on their probability of connection. By incorporating the Hilbert-Schmidt independence criterion (HSIC) as an additional penalty term in the loss we optimize, we ensure that the structure of the latent space is independent of continuous variables which are related to the sampling process. We apply our method to the Spipoll dataset, a citizen science monitoring program of plant-pollinator interactions to which many observers contribute, making it prone to sampling bias because observers are subject to a learning phenomenon as they participate. Finally, we predict changes in the structure of the pollination network in response to variation in landscape composition, with or without taking into account the experience of the observers. The results highlight the importance of correcting for sampling bias, for example the network connectivity greatly increases in agricultural landscapes dominated by livestock when sampling bias is corrected.

**Keywords.** Graphs and networks, Neural networks, Sampling, Applied statistics in ecology

# 1 Introduction

Graph embedding regroups different methods, allowing to represent a network into a vector space in order to gain understanding of key network features. These methods are especially important in the context of large networks. Recently developed graph neural networks (GNNs) enable graph embedding with large-scale methods such as graph isomorphism network, graph attention network or the variational graph variational auto-encoder [Kipf and Welling, 2016]. All these methods can also handle numerous covariates on nodes. GNNs are currently growing in popularity in various domains such as bioinformatics, chemistry and geophysics, but they remain mostly unknown in other research fields.

In ecology, networks have been analyzed to study various types of ecological interactions among species, such as plant-pollinator, predator-prey or host-parasite interactions. The stochastic block model and the latent block model for bipartite graphs are notorious models using latent variable in ecology. While graph embedding methods start being used for ecological networks, GNNs have yet to be diffused in that research field. GNNs could be particularly relevant for ecological networks because very large data sets of interactions among species are now becoming available (e.g. through the development of citizen science programs) in addition to many covariates at the node level (e.g. species name and traits, environmental conditions at the time of interaction observation). An important issue with the analysis of ecological networks concerns the strong effects of sampling effort and methods on the observed network structure [Doré et al., 2021]. One could wish to have an embedding which is independent of a certain set of covariates linked to such sampling effects and related bias. This can be of particular interest for citizen science programs, where biases can arise from the large number observers involved with various experience levels [Jiguet, 2009, Deguines et al., 2018]. It is also known that the biotic and abiotic context of individual plants can

influence pollinator foraging behavior [Arroyo-Correa et al., 2021]. Especially, the land use plays a major role in the structure of the interaction network, as different pollinator groups, such as bees, flies or butterflies, have different affinities for different land-uses, e.g. urban or agricultural areas [Deguines et al., 2012]. For instance, diverse preferences among pollinators for various land-uses might result in reduced interconnectedness of plant-pollinator networks in urban environments compared to rural areas [Geslin et al., 2013, Cortina et al., 2022].

We aim to study the influence of environmental variables, such as the land use on the network structure using GNNs. To explore the impact of environmental conditions on the network structure, we first adapt the graph variational auto-encoder [Kipf and Welling, 2016] to the bipartite case, where the embedding should also be independent of covariates linked to sampling effect. After the learning phase, we analyze how the network embedding predicted by the GNN evolves with changes in input covariates, which should reflect realistic landscape composition.

In the following, a background on GNNs and the HSIC criterion is provided. Then, the model is introduced and applied to the Spipoll data set [Deguines et al., 2012], a citizen science program monitoring plant-pollinator interactions across France since 2010. Finally, we rely on the fitted model to assess how the land use impact the network connectivity.

# 2 Model

Embedding the nodes of a graph in a vector space using a variational graph auto-encoder (VGAE) [Kipf and Welling, 2016] would yield a Gaussian latent representation $Z$. We aim to have $Z$ independent of a set of covariates linked to the sampling process $S$. As we cannot guarantee that $Z$ and $S$ are jointly Gaussian, we will use another criterion than the covariance to have independence between $Z$ and $S$: the Hilbert-Schmidt independence criterion (HSIC), first proposed by Gretton et al. [2005] which is a metric testing for the independence of two variables. Compared to the other proposed methods of embedding, the probabilistic setting of the GVAE fits well with the use of the HSIC, and its generative aspect allows network generation for various ecological contexts.

## 2.1 Bipartite variational graph auto-encoder

We adapt the variational graph auto-encoder from Kipf and Welling [2016] to the bipartite case by considering two graph convolutional networks (GCN), one for each node types.

We consider a biadjacency matrix $B_{i,j}$ of size $n_1 \times n_2$ representing our graph. Let

$$D_1 = diag\left(\sum_{j=1}^{n_2} B_{i,j}\right) \; D_2 = diag\left(\sum_{i=1}^{n_1} B_{i,j}\right)$$

be respectively the row and the column degree matrices. For each $i$ and each $j$ we consider the stochastic latent variables $Z_{1i}$ and $Z_{2j}$ which are described respectively by a $n_1 \times D$ and a $n_2 \times D$ matrices (they share the same number of columns). $X_1$ is a $n_1 \times d_1$ matrix of node

features for the first category, and $X_2$ is a $n_2 \times d_2$ matrix of node features for the second. Finally, we consider the normalized biadjacency matrix $\tilde{B} = D_1^{-\frac{1}{2}} B D_2^{-\frac{1}{2}}$.

### 2.1.1 Encoder

The encoder is defined as

$$q(Z_1, Z_2 | X_1, X_2, B) = \prod_{i=1}^{n_1} q_1(z_{1i} | X_1, B) \prod_{j=1}^{n_2} q_2(z_{2j} | X_2, B)$$

with

$$q_v(z_{vi} | X_v, B) = \mathcal{N}(\mu_{v,i}, diag(\sigma_{v,i}^2)),\ v \in \{1, 2\}$$

with $\mu_v \in \mathbb{R}^d$ and $log(\sigma_v)$ obtained by the $\text{GCN}_v$ defined similarly as in Kipf and Welling [2016]:

$$\text{GCN}_1(X_1, B) = \tilde{B}\text{ReLU}(\tilde{B}^\top X_1 W_{1,1}) W_{1,2}, \quad \text{GCN}_2(X_2, B) = \tilde{B}^\top \text{ReLU}(\tilde{B} X_2 W_{2,1}) W_{2,2}$$

with weight matrices $W_v$. $\text{GCN}_{\mu_v}(X_v, B)$ and $\text{GCN}_{\sigma_v}(X_v, B)$ share the first-layer parameters $W_{v,1}$ and $\text{ReLU}(x) = max(x, 0)$. The parameters $\mu_{1,i}$ and $\mu_{2,j}$ share the same dimension $d$.

### 2.1.2 Decoder

The decoder is defined as

$$p(B | Z_1, Z_2) = \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} p(B_{i,j} | z_{1i}, z_{2j}), \qquad \text{with } p(B_{i,j} | z_{1i}, z_{2j}) = e^{-\frac{||z_{1i} - z_{2j}||^2}{2\sigma^2}}.$$

In the following, we fix $\sigma^2 = 1$.

The full auto-encoder can be summarized as

$$B, X_1, X_2 \xrightarrow[encoder]{q(Z_1, Z_2 | X_1, X_2, B)} Z_1, Z_2 \xrightarrow[decoder]{p(B | Z_1, Z_2)} \widehat{B}.$$

## 2.2 Bipartite and fair graph variational auto-encoder

Our goal is to obtain the latent representation $Z_1$ of the bipartite variational auto-encoder, independent of a protected variable denoted by $S$. To do so, we add the HSIC [Gretton et al.,

2005] computed between the posterior means $\mu_1$ and the protected variable $S$ as a penalty term in the loss:

$$L = \mathbb{E}_{q(Z_1,Z_2|X_1,X_2,B)}[\log p(B|Z_1,Z_2)] - KL[q_1(Z_1|X_1,B)||p_1(Z_1)] \\ - KL[q_2(Z_2|X_2,B)||p_2(Z_2)] + \delta RFF\ HSIC(\mu_1,S) \quad (1)$$

where $p_1, p_2$ are Gaussian priors for $Z_1$ and $Z_2$, $KL$ is the Kullback-Leibler divergence, $\delta$ is a hyperparameter and RFF HSIC is the random Fourier feature [Rahimi and Recht, 2007, Zhang et al., 2018] estimation of the HSIC. At the end of the learning, we can compute the p-value of the HSIC test [Gretton et al., 2007] to check that the optimization of the loss has made the latent representation $Z_1 \sim \mathcal{N}(\mu_1, diag(\sigma_1^2))$ independent of $S$. Note that adding the HSIC in the loss may deteriorate the reconstruction of the original data through the decoder since an independence constraint has been added.Full description of the model is available in Anakok et al. [2024].

# 3 Application on Spipoll data set

## 3.1 Context

We apply the proposed method on the Spipoll [Deguines et al., 2012] data set, a French citizen science program which aims to monitor plant-pollinator interactions across metropolitan France since 2010. This monitoring follows a simple protocol: briefly, volunteers can choose a flowering plant where and when they like, and during 20 minutes, take pictures of all different insects that land on the flowers of the monitored plant. Then, using an online identification tool, they identify each different insect that has been photographed and upload their data on a dedicated website. Each participation is thus a set of insect interactions with a given plant species that have been observed at a given time and place, and by a given volunteer whose specific skills could affect the quality of the observation. The date and place of observations allowed us to extract corresponding climatic conditions as covariates, from the European Copernicus Climate data set, and the corresponding land use proportion with the Corine Land Cover (CLC).

A common practice in ecology to study plant-pollinator interactions is to consider plant and insect species as nodes of a bipartite network, with edges determined by the observations of insects pollinating plants. Since our data are at the participation level (a session of observation corresponds to a plant species with all the observed insects on that plant during 20 minutes), we consider a bipartite network where the first type of nodes are session of observations, and the second type are insect species observed during the session. Each session has the previously mentioned covariates and a one-hot encoding describing the plant genus. Link prediction task in this situation aims to predict which insect will be present during a given observation session. However, we still wish to ultimately obtain a bipartite plant-insect network, which is standard in this field of study. To ensure that the latent space could also be used to create a plant-insect network, we propose to draw for each taxon of plant one observation from the set of all the session where this plant was monitored.This would

generate another latent space corresponding to plant and insect species, and using the same decoder, would generate a plant-insect network. More details about this model are available in the full presentation of our model [Anakok et al., 2024].

We fit our model, taking into account the specific requirements of the Spipoll dataset. We consider the observation period of the Spipoll dataset from 2010 to 2020 included in metropolitan France, on a set of 83 plant genus that have been monitored every year. This lead to 26267 observation sessions during which 306 insect taxa have been observed. The observation session-insect matrix has a total of 94 909 plant-pollinator interactions reported, and the plant-insect matrix has 9 754 different interactions. The covariates related to the observations sessions are $X_1 = (P, t, \Delta_T, CLC)$ where $P$ is a binarized categorical variable (83 columns) giving the plant genus, $t$ contains the day and the year of observation, $\Delta_T$ is the difference between the average temperature on the day of observation and the average of temperatures measured from 1950 to 2010 at the same observation location and $CLC$ describes the proportion of land use with 44 categories (see fig. 1) in a 1000m radius around the observation location. We only consider covariates for the observation sessions, thus the covariates for insects are set as $X_2 = I_{n_2}$. While citizen science programs facilitate the accumulation of observed data, the sampled data may be biased by the participating observers. To take into account this bias, we propose in our model to define $S$, the protected variable, as the number of participation from the user at the time of observation. This number of participation would work as a proxy of the user's experience. By employing this measure, we aim to construct a latent space that remains unaffected by variations in observers' experience levels.

## 3.2 Results : Influence of the landscape on network connectivity

Once the model has been fit, we target the impact of the environmental variables on the network structure in particular we focus on the composition of the landscape described by the CLC. To do so, we could change the covariates input related to the landscape to create "pure" landscapes to assess its effect on the predicted network. However, setting a pure landscape with 100% for a specific type and the rest at 0% (e.g. 100% continuous urban fabric and the rest at 0%) would yield unrealistic landscapes. Moreover, the transition from one type of landscape to another is also not simple to simulate. In order to get more realistic landscape simulation, we can seek for typical landscapes by performing a clustering on the CLC indiced. Since the CLC are compositional data (proportions of land use), they are transformed through an isometric log ratio transformation (ILR) and a principal component analysis is computed as proposed by Aitchison [1983]. The typical landscapes are obtained as the centroids of a k-means clustering on the first component of the PCA. Such a clustering is displayed in fig. 2. The centroids of the k-means algorithm are represented with pie-charts. The pie charts composition are detailed in fig. 1, and we assume that they represent typical landscape at places of sampling. For example in fig. 1, landscape 5 is mostly composed of discontinuous urban fabric, which is typical of sampling performed in metropolitan area. Landscape 3 is mostly composed of pastures and forest, landscapes 1 and 2 are made of culture and forest, with a bit of urban fabric, and landscape 4 is mostly arable land. The ILR transform will also allow us to simulate realistic landscape proportion transition, from

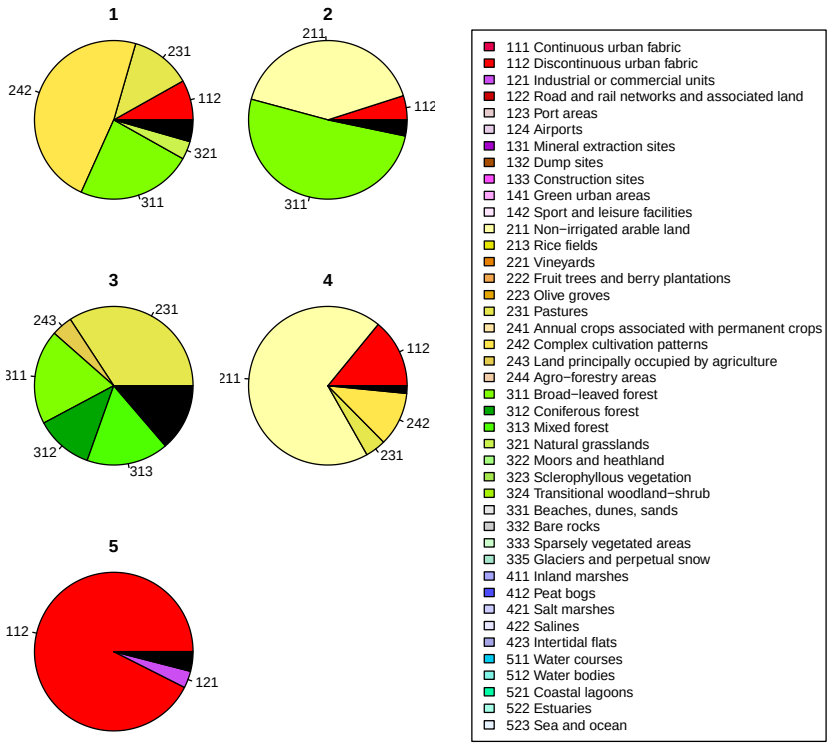



Figure 1: Typical landscapes at sampling localization and CLC legends. All proportions smaller than 4% are regrouped and colored in black for better visibility.

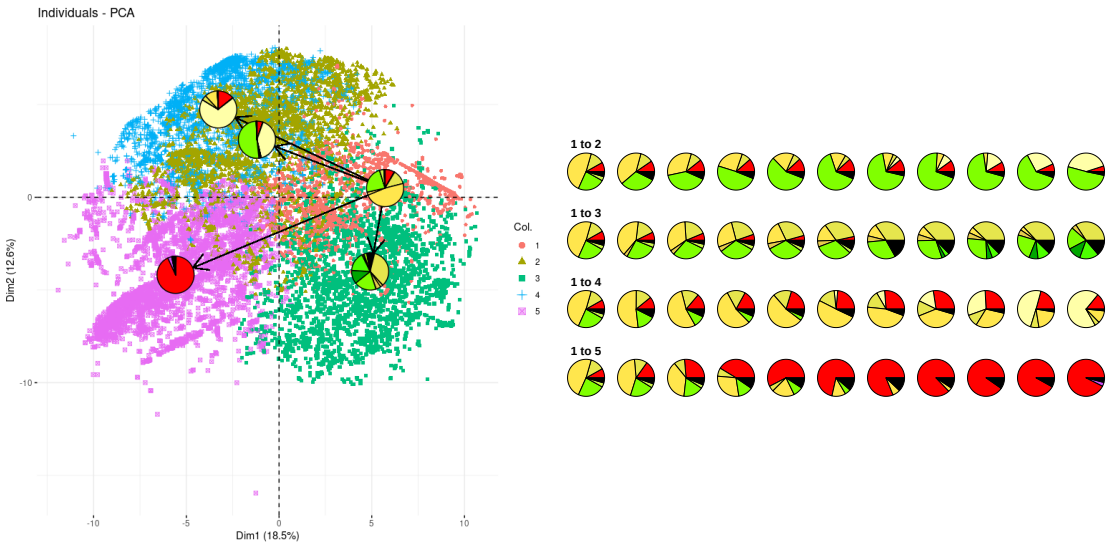



Figure 2: On the left, the first two principal components of the PCA performed on the ILR of the land use proportion of each in observation session in the Spipoll dataset. The arrows represent some simulated trajectories of landscape change from the first typical landscape to the others. On the right, we see how these trajectories change the simulated landscape.

one landscape to the other (fig. 2) by using convex combination of the ILR transform of the typical landscapes.

We fit the model 100 times on the Spipoll dataset,with different initializations and different train-test splits, with and without taking into account the observers' experience levels to compare the predictions. After the learning phase, we input the simulated transition of landscape in the fitted algorithm and observe how it changes the network. The result for connectivity prediction can be seen in fig. 3. Taking into account the observers' experience level increases the probability of connection. Going from landscape 1, which is mostly composed of complex cultivation patterns and broad-leaved forest, to landscape 3, which is mostly made of pasture and forest, drastically decreases the connectivity of the plant-pollinator network from 0.4 to almost 0 (trajectory 1_3), unless you take into account the observers' experience levels (trajectory F 1_3), where we see that the decreasing is less pronounced. The uncorrected connectivity seems to be the highest in landscape 1 but with the correction, the connectivity in landscape 4 is on par with the one in landscape 1. The decline in complex cultivation patterns from landscape 1 to other landscapes also correlates with a decline in connectivity in the network when we do not take into account the observers' experience levels, but this effect seems to be less visible in the debiased estimation. These observations provide insights into the relationship between land use and the network structure, and how taking into account the sampling bias could change our understanding of the network.

# 4 Perspectives

In this exploratory work, we demonstrate how our approach can take into account the bias induced by the participating observers, alongside an examination of how varying covariates induces changes in the plant-pollinator network structure. However, there remains further exploration into the impact of these environmental covariates. Additional metrics, such as modularity, nestedness or robustness can be studied to enhance the understanding of the network evolution. Categorizing organisms by order of insect or plants could reveal instances where overall connectivity decreases for certain order while increasing for others. Expanding the study to the effect of the temperature, or other environmental covariates could also provide valuable insights on various ecological questions. Other sampling bias, such as the uneven distribution of the sampling location on the territory, could be taken into account. All of these aspects will be taken into consideration in future work.

# References

J. Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983. ISSN 00063444. URL http://www.jstor.org/stable/2335943.

E. Anakok, P. Barbillon, C. Fontaine, and E. Thebault. Bipartite graph variational auto-encoder with fair latent representation to account for sampling bias in ecological networks, 2024.

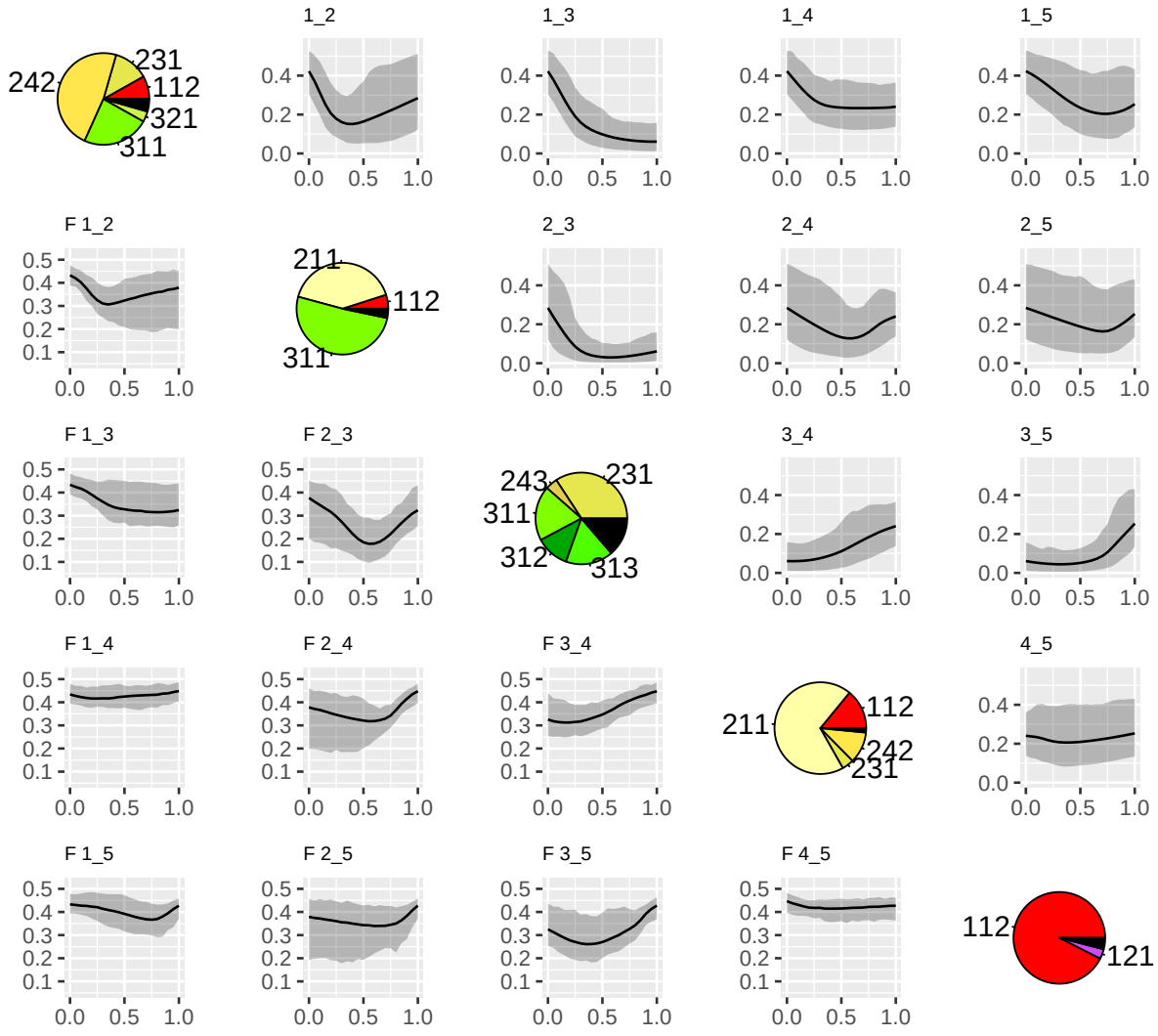



Figure 3: Evolution of network connectivity along the trajectories of landscape. The trajectory is simulated from typical landscape $i$ to $j$ (written as "$i\_j$"). Above the diagonal, we can see the evolution of the predicted network connectivity without taking account of the observers' experience levels (uncorrected connectivity). Below the diagonal, the prediction takes into account the observers' experience levels (corrected connectivity, noted with a "F"). The simulated trajectories of the first row, and by symmetry the first column, are represented in fig. 2. The mean of the results is represented with a black curve, and 95% of the predictions are in the gray area.

B. Arroyo-Correa, I. Bartomeus, and P. Jordano. Individual-based plant–pollinator networks are structured by phenotypic and microsite plant traits. *Journal of Ecology*, 109(8):2832–2844, 2021. ISSN 1365-2745. doi: 10.1111/1365-2745.13694.

C. A. Cortina, J. L. Neff, and S. Jha. Historic and Contemporary Land Use Shape Plant-Pollinator Networks and Community Composition. *Frontiers in Ecology and Evolution*, 10, 2022. ISSN 2296-701X.

N. Deguines, R. Julliard, M. de Flores, and C. Fontaine. The Whereabouts of Flower Visitors: Contrasting Land-Use Preferences Revealed by a Country-Wide Survey Based on Citizen Science. *PLOS ONE*, 7(9):e45822, Sept. 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0045822.

N. Deguines, M. De Flores, G. Loïs, R. Julliard, and C. Fontaine. Fostering close encounters of the entomological kind. *Frontiers in Ecology and the Environment*, 16(4):202–203, May 2018. ISSN 15409295. doi: 10.1002/fee.1795.

M. Doré, C. Fontaine, and E. Thébault. Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale. *Global Change Biology*, 27(6):1266–1280, 2021. ISSN 1365-2486. doi: 10.1111/gcb.15474. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15474.

B. Geslin, B. Gauzens, E. Thébault, and I. Dajoz. Plant Pollinator Networks along a Gradient of Urbanisation. *PLOS ONE*, 8(5):e63421, May 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0063421.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 63–77, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31696-1. doi: 10.1007/11564089_7.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

F. Jiguet. Method learning caused a first-time observer effect in a newly started breeding bird survey. *Bird Study*, 56(2):253–258, July 2009. ISSN 0006-3657, 1944-6705. doi: 10.1080/00063650902791991.

T. N. Kipf and M. Welling. Variational graph auto-encoders. 2016. doi: 10.48550/ARXIV.1611.07308. URL https://arxiv.org/abs/1611.07308.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20, Jan. 2007.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, Jan. 2018. ISSN 1573-1375. doi: 10.1007/s11222-016-9721-7.

# Analyse statistique d'interventions pour la réduction de la consommation d'eau potable

Charlotte Rous [1] & Julia Barrault [2] & Vincent Couallier [3] & Laetitia Couderc [4] & Charlotte Sakarovitch [5]

[1] *SUEZ Le LyRE, France et charlotte.rous@suez.com*
[2] *SUEZ Le LyRE, France et julia.barrault@suez.com*
[3] *IMB, CNRS, Université de Bordeaux, France et vincent.couallier@u-bordeaux.fr*
[4] *SUEZ Le LyRE, France et laetitia.couderc@suez.com*
[5] *SUEZ Le LyRE, France et charlotte.sakarovitch@suez.com*

**Résumé.** La préservation de la ressource en eau est actuellement un enjeu clé en France. Cet enjeu deviendra chaque année de plus en plus central et prégnant notamment en raison du réchauffement climatique et de l'évolution démographique. Dans ce contexte, la compréhension des mécanismes liés à la consommation d'eau est nécessaire afin de déclencher des changements d'habitudes garantissant à terme une réduction de la consommation. Dans le cadre d'une étude à grande échelle sur la Communauté d'Agglomération du Bassin de Brive entre 2022 et 2025, la réduction de la consommation d'eau des usagers particuliers est évaluée en s'appuyant sur les sciences comportementales. Pour cela, nous nous basons sur les données des compteurs connectés (télérelève), fournies par Suez depuis mars 2022 pour les premières mesures, afin d'évaluer différentes interventions (action de communication ou distribution de dispositifs d'économie d'eau), chacune ayant pour but commun une réduction de la consommation d'eau domestique. Pour ce faire, chaque intervention a été mise en place à la façon d'un essai randomisé contrôlé. Après un pré-traitement des données, nous proposons d'utiliser les modèles linéaires mixtes prenant en compte les données répétées et les aspects pré/post intervention afin d'évaluer l'effet de chaque intervention. L'analyse est en cours de réalisation, les premiers résultats sont attendus en avril.

**Mots-clés.** Essai Randomisé Contrôlé ; Maîtrise des consommation et économies d'eau ; Modèle Linéaire Mixte

**Abstract.** The preservation of water resources is a current issue as well as a future concern in France, particularly due to global warming and demographic changes. In this context, understanding the mechanisms related to water consumption is necessary in order to trigger changes in habits that will ultimately guarantee a reduction in resources consumption. As part of a large-scale study on the Agglomeration Community of the Brive Basin between the years 2022 and 2025, the reduction in water consumption by individual users is evaluated based on behavioral sciences. To complete this task, we rely on data obtained from connected water meters (telemetering), provided by Suez since March 2022 for the first measurements. The data gathered from these devices is used to analyze and evaluate different interventions (communication action or distribution of water-saving devices), each with the common goal of reducing domestic water consumption. To that end, each intervention was implemented in the manner of a randomized controlled trial. After pre-processing the data, we propose to use

Linear Mixed Models taking into account repeated data and pre/post intervention aspects to evaluate the effect of each intervention. The analysis is currently being carried out, with the first results expected in April.

**Keywords.** Randomised Controlled Trial ; Water savings ; Linear Mixed Model

# 1 Introduction

La Communauté d'Agglomération du Bassin de Brive, particulièrement touchée par la sécheresse intense de 2019, a pour ambition de préserver la ressource en eau de son territoire. L'objectif de ce territoire est d'atteindre une réduction des prélèvements sur la ressource de 21% en 7 ans, en partie grâce à une réduction de 8% de la consommation d'eau après compteur, c'est-à-dire l'eau réellement consommée par les usagers du territoire. Le projet DEM'EAU (*DEM*onstrateur d'économie d'*EAU*) s'inscrit dans cette trajectoire. Ce projet cible uniquement les usagers particuliers du service de l'eau et a pour but de comparer et quantifier l'efficacité de différentes interventions sur la réduction des consommations. Jusqu'à récemment, la consommation des ménages était évaluée de manière annuelle grâce à des campagnes de relèves à pied permettant d'établir les factures des usagers. La mesure d'une réduction très faible de la consommation d'eau est difficilement perceptible sur une consommation annuelle. Sur l'Agglomération de Brive un nouvel outil de télérelève journalière se déploie progressivement (et à terme exhaustivement) et nous permet une mesure plus efficace et plus fiable de par sa fréquence d'acquisition de données bien plus grande qu'un simple index annuel.

Afin d'identifier les moyens les plus efficaces pour préserver la ressource en eau, 4 interventions visant à réduire les consommations des ménages ont été mises en place pour être évaluées et comparées à l'aide d'essais randomisés contrôlés :

— **Intervention 1** : Distribution de nudges par voie postale : Thaler et Sustein [2008], définisse le nudge (ou coup de pouce) comme instrument simple à mettre en place et relativement peu coûteux qui permet d'aider les individus à prendre des décisions bénéfiques pour eux-mêmes ou pour la société tout en leur laissant la liberté de choix. Dans cette intervention, il s'agit d'un livret montrant l'intérêt de la préservation de la ressource accompagné de stickers incitant aux éco-gestes (douche, arrosage des fleurs avec de l'eau réutilisée).

— **Intervention 2** : Distribution d'un pommeau de douche hydro-économe aux personnes se présentant en mairie suite à un courrier électronique ou postal les informant de cette mise à disposition. Les pommeaux étaient disponibles 3 semaines après envoi du courrier. De marque *Hydrao*© [1], ces pommeaux, en plus de réduire le débit de la douche, éclairent le jet d'eau avec des couleurs différentes en fonction du volume d'eau consommé et sont connectés en bluetooth à une application sur smartphone permettant de surveiller sa consommation. Le fournisseur annonce un objectif de réduction de 70% pour l'eau consommée sous la douche.

1. https ://www.hydrao.com/fr/notre-concept/fonctionnement

— **Intervention 3** : Installation d'un pommeau de douche hydro-économe à domicile. Il s'agit des même pommeaux que dans l'intervention 2. Pour cette intervention, les foyers sont contactés par téléphone pour leur proposer un rendez-vous pour la pose du pommeau à domicile. Il s'agit ici d'évaluer une autre politique de mise à disposition que celle de l'intervention 2.
— **Intervention 4** : Mise à disposition d'une application numérique, nommée *ModeEco*. Cette application, liée aux données télérelevées, permet à l'usager de connaître en temps réel sa consommation d'eau et d'obtenir également des conseils en vue de diminuer sa consommation (alertes fuites et surconsommation, comparaison de la consommation du ménage à un référentiel, etc.)

# 2 Population et protocole d'étude

L'étude a lieu sur le territoire de la Communauté d'Agglomération du Bassin de Brive composée de 38 communes classées en 3 catégories (urbain, péri-urbain, ou rural, figure 1), regroupant au total 44 815 ménages qui seront à terme tous dotés de compteurs connectés.

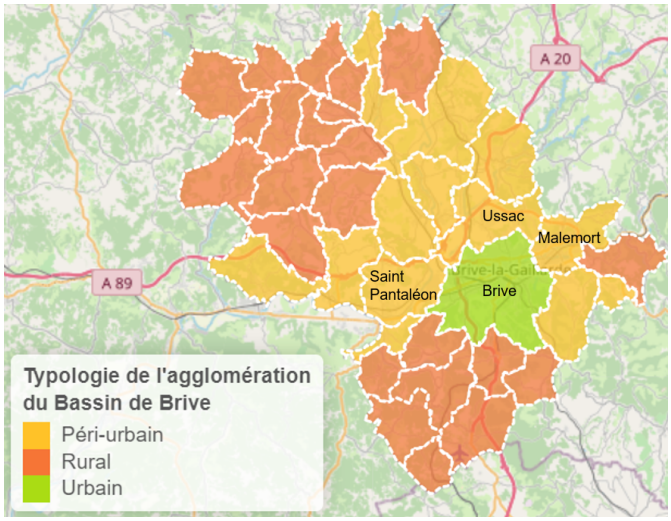



FIGURE 1 – Périmètre de l'étude

L'évaluation des trois premières interventions repose sur un protocole de type essai randomisé contrôlé, avec un groupe de ménages touchés par l'intervention et sélectionnés aléatoirement dans la population éligible, le reste formant le groupe contrôle. Les interventions 2 et 3 sont destinées aux 6 000 ménages les plus gros consommateurs d'eau. L'intervention 1 est destinée aux ménages restants. Pour des raisons matérielles d'activation de l'application *ModeEco* par commune entière, l'intervention 4 nécessite un tirage aléatoire de communes, tous les ménages des communes sélectionnées intègrent donc le groupe test (échantillonage par grappe).

Afin de prendre en compte les variations saisonnières ou infra saisonnières dues notamment

à la météo, aux périodes de vacances scolaires ou aux évènements touristiques, les données de consommations de chaque groupe sont collectées avant et après l'intervention. Un modèle mixte permettra de prendre en compte l'aspect répété des données longitudinales de consommations étudiées (Verbeke [1997]).

Les périodes de mise en place ainsi que les dates d'évaluation de ces interventions sont détaillées dans la figure 2. L'expérimentation prévue pour durer 2 ans doit prendre fin au premier trimestre 2025.

Grâce à l'expertise en science sociale du LyRE et aux études déjà réalisées dans ce domaine par l'équipe projet sur des échantillons plus réduits (Lecourt et al. [2023]), des hypothèses d'efficacité des différentes interventions ont pu être posées. Pour le nombre de sujets nécessaires, un calcul naïf de puissance pour la comparaison de deux échantillons gaussiens a pu être mis en place avec des ordres de grandeur réalistes pour les consommations et les dispersions individuelles.

Suite à l'hypothèse d'une réduction de la consommation d'eau de 5% pour les nudges, le groupe test de l'intervention 1 est de 3 100 ménages avec un groupe contrôle de 5 200 ménages. Pour les interventions 2 et 3, l'hypothèse de réduction de la consommation d'eau des pommeaux est de 20%. 1 000 pommeaux ont été envoyés à chacun des groupes test des interventions 2 et 3.

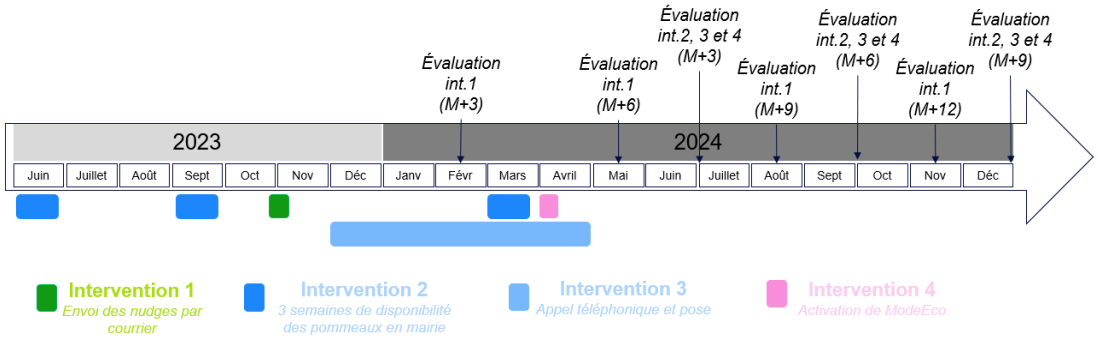



FIGURE 2 – Planning du déploiement des actions et des évaluations

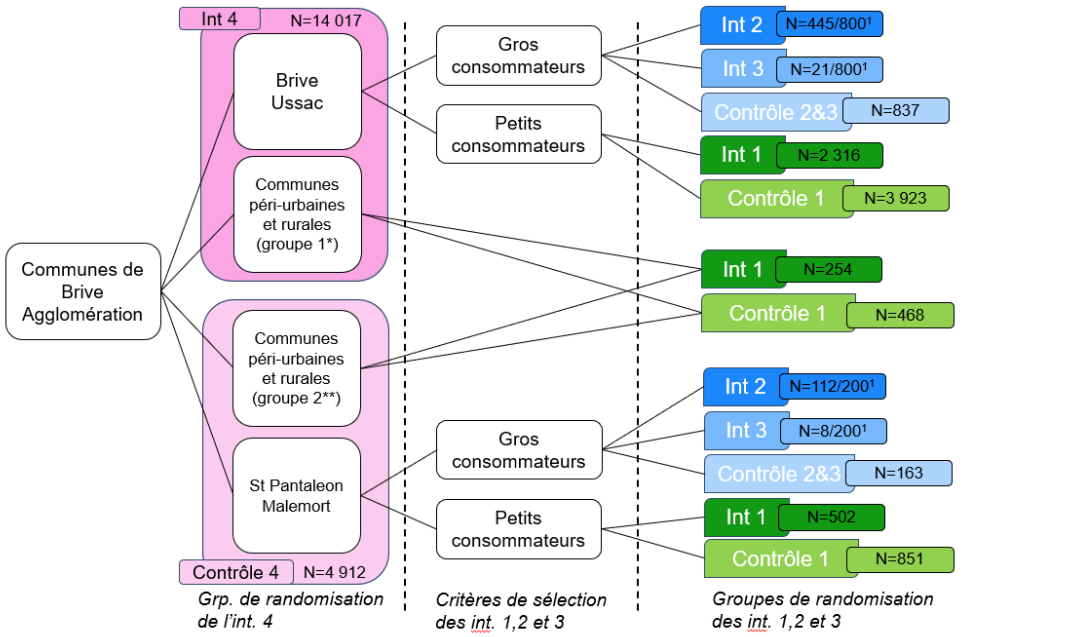
La figure 3 synthétise la répartition des interventions sur les ménages de l'agglomération de Brive disposant de la télérelève.

Comme indiqué sur la figure 3, les ménages de Brive et la moitié des communes péri-urbaines et rurales feront l'objet de l'intervention 4 consistant au déploiement de l'application *ModeEco*. Le reste des ménages de l'agglomération (toutes les communes ne faisant pas partie du groupe test) inclus dans l'étude servira de groupe contrôle à cette intervention. La mise en place de l'application est possible uniquement à l'échelle communale et non pas par usager. Les communes tests ont été choisies en stratifiant sur la typologie de la commune (rural, péri-urbain ou urbain). Ainsi, l'intervention 4 induit un emboîtement des interventions de par son mode d'activation, il faudra alors prendre en compte les potentiels effets dus aux autres interventions lors de l'évaluation de cette action. À l'inverse, il faudra ajuster les autres modèles

avec l'intervention 4 afin d'isoler les effets des autres interventions.

Les interventions 2 et 3, quant à elles, sont réparties sur les ménages de Brive et 3 autres communes péri-urbaines (Malemort, Ussac et Saint Pantaléon).

L'intervention 1 a été répartie sur l'ensemble des communes de l'agglomération et ne concerne que les plus petits consommateurs d'eau afin de ne pas interférer avec les interventions 2 et 3.




¹ Effectif atteint au 30/01/2024 / Effectif cible
*Groupe 1 : Allasac, Chasteaux, Dampniat, Donzenac, Estivals, Mansac, Noailles, Objat, Perpezac le Blanc, Rosiers-de-Juillac, Saint Aulaire, Saint Cyr La Roche, Saint Robert, Saint Solve, Voutezac, Yssandon
**Groupe 2 : Ayen, Brignac la Plaine, Chartière Ferrière, Cublac, Jugeals Nazareth, Larche, Lissac sur Couze, Louignac, Nespouls, Saint-Bonnet-la-Rivière, Segonzac, Saint Cernin de Larche, Saint Cyprien, Saint Viance, Varetz, Vars sur Roseix

FIGURE 3 – Répartition des ménages dans les différentes interventions

# 3 Définition de la variable d'intérêt

Afin d'évaluer la réduction de consommation d'eau des particuliers, nous avons à disposition un index journalier pour chaque ménage de notre étude (données issues de la télérelève). Cependant, de par la présence de données manquantes (échec de transmission), de la très grande variabilité inter-journalière de la consommation d'eau, du manque d'information sur la présence des usagers jour par jour, de l'échelle temporelle des effets attendus, nous proposons une définition adaptée de la consommation d'eau journalière moyenne mensuelle. Un premier filtre des données a été réalisé en éliminant les jours où la consommation est supérieure à 2000 litres (consommation exceptionnelle correspondant à la présence de plus de 13 personnes dans le foyer ou encore une consommation aberrante liée à une erreur de mesure), et en ne conservant que les jours de présence, identifiés par une consommation journalière

mesurée supérieure à 5 litres, qui correspond dans la littérature à la consommation d'une une seule chasse d'eau. On note l'indicatrice de prise en compte du jour $d$ dans notre étude $p(d) = \mathbb{1}_{\text{Conso(d)}\geq 5} \times \mathbb{1}_{\text{Conso(d)}\leq 2000}$, et on définit la consommation journalière au jour $d$ par $\text{Conso(d)} = [\text{Index}(d)] - [\text{Index}(d-1)]$. La consommation journalière mensuelle au mois $m$ est alors définie par :

$$\text{Conso(m)} = \frac{\sum_{d\in m} \text{Conso(d)} \times p(d)}{\sum_{d\in m} p(d)}.$$

La variable d'intérêt de notre étude est le logarithme de la consommation journalière (en litres) $\text{LogConso}(m) = \ln(\text{Conso}(m))$, le passage au logarithme permettant une interprétation en pourcentage de la réduction de la consommation.

# 4 Évaluation

Dans le cadre des évaluations d'interventions, il est d'usage de distinguer l'évaluation en intention de traiter (ITT) ou en *per procole* (PP). L'ITT compare les résultats des groupes test et contrôle indépendamment de la réelle utilisation des dispositifs par les ménages dans le groupe testé. Dans l'analyse per protocole, seuls les individus du groupe testé qui auront utilisé les dispositifs seront pris en compte (Lachin, J. M., 2007, McNamee, R., 2009, Sedgwick, P., 2015).

Pour évaluer nos différents dispositifs, il est naturel de considérer des modèles linéaires mixtes, adaptés aux mesures répétées sur le sujet "ménage". L'évaluation de l'ensemble de nos actions requiert l'ajustement de plusieurs modèles puisque la population des groupes contrôle est différente pour chacune des interventions (gros consommateurs pour les interventions 2 et 3, petits consommateurs pour l'intervention 1). De plus, l'intervention 4 déployée par commune va cibler des ménages ayant pu être concernés par une autre intervention. Il faudra donc prendre cela en compte dans les différents modèles afin d'isoler l'effet recherché.

Pour analyser l'impact de l'intervention 1 (population en vert sur la figure 3), le modèle suivant est proposé :

$$\begin{aligned}\text{LogConso(im)} = \beta_0 + \beta_1 \mathbb{1}_{Int.1}(\text{im}) + \beta_4 \mathbb{1}_{Int.4}(\text{im}) + \beta_5 \text{mois}_{cos}(\text{m}) + \beta_6 \text{mois}_{sin}(\text{m}) + \\ + \beta_7 \text{Commune(i)} + \gamma_i + \epsilon(\text{im})\end{aligned}$$

où $i$ correspond au ménage, $\beta_1$, $\beta_4$ correspondent à l'effet de l'intervention 1 et 4, $\beta_5$, $\beta_6$, l'ajustement sur le mois à l'aide de transformations trigonométriques, $\beta_7$, l'ajustement sur la commune, $\gamma_i$, l'effet aléatoire sur le ménage et $\epsilon(\text{im})$ le terme d'erreur. Dans ce modèle, $\beta_4$, ne sera pas interprété, c'est le coefficient associé à l'intervention 4 sur laquelle on se doit d'ajuster. En effet, l'ajustement sur l'intervention 4 est nécessaire puisque l'application va être communiquée sur différentes communes, mais pas à l'ensemble des ménages de notre panel (intervention 1), il permettra de séparer l'effet de ces deux interventions.

Pour analyser l'impact des interventions 2 et 3 (le groupe contrôle étant identique, en bleu

sur la figure 3), le modèle suivant est proposé :

$$\begin{aligned}\text{LogConso(im)} = \beta_0 + \beta_2 \mathbb{1}_{Int.2}(\text{im}) + \beta_3 \mathbb{1}_{Int.3}(\text{im}) + \beta_4 \mathbb{1}_{Int.4}(\text{im}) + \beta_5 \text{mois}_{cos}(\text{m}) + \\ \beta_6 \text{mois}_{sin}(\text{m}) + \beta_7 \text{Commune(i)} + \gamma_i + \epsilon(\text{im})\end{aligned}$$

où $\beta_2$, $\beta_3$ correspondent aux effets des interventions 2 et 3. Dans le même objectif que précédemment, le $\beta_4$ ne sera pas interprété, afin d'ajuster sur l'intervention 4.

Pour analyser l'impact de l'intervention 4 (les encadrés rose sur la figure 3), le modèle suivant est proposé :

$$\begin{aligned}\text{LogConso(im)} = \beta_0 + \beta_1 \mathbb{1}_{Int.1}(\text{im}) + \beta_2 \mathbb{1}_{Int.2}(\text{im}) + \beta_3 \mathbb{1}_{Int.3}(\text{im}) + \beta_4 \mathbb{1}_{Int.4}(\text{im}) + \\ \beta_5 \text{mois}_{cos}(\text{m}) + \beta_6 \text{mois}_{sin}(\text{m}) + \beta_7 \text{Commune(i)} + \\ \gamma_i + \epsilon(\text{im})\end{aligned}$$

Dans ce modèle, l'intérêt porte seulement sur l'interprétation du $\beta_4$. Les autres paramètres servent d'ajustement.

# 5 Résultats

À ce jour il n'y a pas encore de résultats. Comme illustré dans le planning des évaluations (figure 2), l'analyse de l'effet de l'intervention 1 après 3 mois est en cours de réalisation et les résultats pourront être communiqués en avril. Une précédente étude sur Bordeaux Métropole (Lecourt et al. [2023]), sur un échantillon plus petit (630 ménages), nous laisse penser que les effets des nudges (intervention 1) sont réels mais de faible envergure. Ceci concorde avec les enseignements de la littérature sur des expérimentations similaires menées à l'étranger par Nisa et al. [2019].

Il est à noter que les dates d'inclusion (mise en place du compteur connecté) ainsi que les dates d'intervention pour les panels 2 et 3 sont différentes pour chaque ménage, c'est pourquoi une analyse prenant en compte l'aspect longitudinal des données comme le modèle mixte est indispensable. A terme ce projet permettra d'évaluer les effets dans le temps des différentes intervention de réduction de la consommation en eau grâce à des analyses réalisées à $M_0 + 3$, $M_0 + 6$, $M_0 + 12$ et $M_0 + 24$ (où $M_0$ correspond au mois de l'intervention).

# Bibliographie

Bates, D. (2010). Ime4 : Mixed-effects modeling with R. Récupéré sur http ://lme4.r-forge.r-project.org/book/

Bates, D., Mächler, M., & Bolker, B. (2015). Fitting Linear Mixed-Effects Models Using Ime4. Journal of Statistical Software, 67(1), pp. 1-48.

Jiang, J., & Nguyen, T. (2007). Linear and generalized linear mixed models and their applications. New York : Springer, 1.

Lachin, J. M. (2007). Intention-to-Treat Analysis. Wiley Encyclopedia of Clinical Trials, 1-9.

Lecourt, M., et al. (2023) Les sciences comportementales au service de la sobriété des consommations d'eau. Communications ASFEE (Conference of the French Association of Experimental Economics), et ASTEE (Association scientifique et technique pour l'eau et l'environnement)

Lecourt, M., et al. (2023) Les sciences comportementales au service de la sobriété des consommations d'eau. TSM à paraître (Techniques Sciences et Méthodes)

McNamee, R. (2009). Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. Statistics in medicine, 28(21), 2639-2652.

Nisa, C., Bélanger, J., & Schumpe, B. e. (2013). Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. Nat Commun, 4545(10).

Sedgwick, P. (2015). Intention to treat analysis versus per protocol analysis of trial data. Bmj, 350.

Thaler, R., & Sustein, C. (2008). Nudge. Improving Decisions About Health, Wealth, and Happiness. (Y. U. Press, Éd.) New Haven.

Verbeke, G. (1997). Linear Mixed Models for Longitudinal Data. In : Linear Mixed Models in Practice. Lecture Notes in Statistics, vol 126. Springer, New York, NY. https ://doi.org/10.1007/978-1-4612-2294-1_3

# Extremiles in environmental research and meteorology

Abdelaati Daouia [1] & Thibault Laurent [2]

[1] *Toulouse School of Economics, France, Abdelaati.Daouia@tse-fr.eu*
[2] *Toulouse School of Economics, CNRS, France, Thibault.Laurent@tse-fr.eu*

**Résumé.** La question du changement climatique s'est imposée comme un enjeu politique mondial. Records de chaleur, multiplication des catastrophes météo, fonte des glaces, déclin de la nature : les preuves de l'impact dévastateur des activités humaines sur la planète s'accumulent. De nombreuses études scientifiques s'appuient sur des données météorologiques, devenues plus nombreuses et facilement accessibles. Dans ce travail, nous appliquons des outils statistiques habituellement utilisés dans l'évaluation des risques dans le secteur de la banque et de l'assurance, à des données environnementales mesurant le changement climatique. Plus précisément, nous calculons les mesures de risque extremiles (Daouia *et al.*, 2019, 2022) sur deux échelles géographiques différentes : la donnée satellite elle-même d'une part et à l'échelle du pays d'autre part. Les mesures de risque sont calculées à différente période entre 1980 jusqu'à 2023. Les résultats montrent une tendance globale à la hausse des risques de fortes précipitations et de vagues de chaleur au cours du temps. Par ailleurs, toutes les zones géographiques ne sont pas exposées de la même façon. Les risques de précipitations élevées sont concentrés autour des tropiques alors que les risques de vagues de chaleurs sont très forts dans les régions proches des pôles, en Afrique centrale, le nord de l'Amérique du Sud, en Europe continentale et en Amérique du Nord.

**Mots-clés.** Environnement et statistique, mesures de risques, quantiles, extemiles, données météorologiques.

**Abstract.** Climate change has emerged as a global political issue. Evidence of the devastating impact of human activities on the planet is accumulating, including record heat, increasing number of weather disasters, melting ice, decline of nature. Many scientific studies are based on meteorological data that has become abundant and easily accessible. In this work, we apply a recent statistical tool from risk handling in finance and insurance to environmental data for assessing some aspects of climate change. More specifically, we use the concept of extremiles, which defines an asymmetric least squares analog to quantiles (Daouia *et al.*, 2019, 2022), as an alternative risk measure to the standard Value at Risk. In order to quantify how extreme precipitation and heatwaves have become over time, we perform extremile estimation and inference at different time periods between 1980 and 2023 on two different geographical scales: the satellite data itself besides the country scale. The results show an overall increasing evolution in heavy precipitation and extreme heatwaves. However, the tail exposure differs following the geographical location. The risks of high precipitation are concentrated around the tropics, while the risks of heatwaves are very strong in regions near the poles, in Central Africa, northern South America, Europe, and North America.

**Keywords.** environment and statistics, risk measures, quantile, extremile, climate change data.

# 1 Introduction

Meteorological data provide nowadays alarming aspects of global warming and climate change with tremendous impact on society, as they are among the most prominent topics of discussion. In statistical decision theory, one is typically interested in the analysis of heavy precipitation (Fischer and Knutti, 2016) and extreme temperatures (Katz and Brown, 1992) due to their adverse effects, such as floods or droughts. The risk of disasters is usually summarized by a risk measure that is estimated from historical data. The most common risk measure used in environmental research and meteorology is Value at Risk, which is a simple tail quantile of the underlying distribution. It is, however, known that quantiles only depend on the frequency of tail observations and not on their severity. Therefore, sample quantiles are either insensitive to the magnitude of infrequent catastrophic events, or they completely breakdown for high tail probability levels.

Extremiles (Daouia *et al.*, 2019, 2022) have recently emerged as an alternative least squares analog of quantiles. They are determined by weighted expectations rather than tail probabilities, and are preferred in this respect over quantiles in terms of alertness and reactivity, as well as resistance to outliers. Their use as an instrument of risk protection in finance and actuarial science has revealed their specific merits and strengths in the axiomatic theory of risk measures. This article is the first work to implement their contribution to mitigating the impact of climate extremes. In recent years, meteorological data has become abundant and easily accessible at extremely fine spatial and temporal scales. In Section 2, we construct, at different geographical scales, two annual indicators of climate change: a first indicator linked to heavy precipitation and another one related to climate warming. Section 3 presents the extremile risk measure, while Section 4 provides our preliminary results.

# 2 Data

We use the Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2), provided by the National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO) from 1980 to 2023. MERRA-2 is fully described in Gelaro *et al.* (2017) and used in many recent scientific articles (see, for instance, `https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/pubs/`). All of the MERRA-2 atmospheric variables are provided at $0.5^\circ \times 0.625^\circ$ spatial resolution: It has 576 points in the longitudinal direction and 361 points in the latitudinal direction, *i.e.* a total of $576 \times 361 = 207,936$ measurement cells covering the whole earth.[1] In addition, several time steps are available (hour, day, month, year) depending on the variable considered.

**Heavy precipitation:** In cities and towns, high precipitation rates overwhelm storm drains and cause flash flooding. They can also be cause for great concern in rural areas by drowning crops, eroding topsoil, and damaging roads. We consider as a measure of heavy precipitation the maximum consecutive three-day precipitation in a year, which is also adopted by

[1]Note that the maximum distance between two measurement cells is around $70km$ on the equator line.

the Climate Atlas of Canada (see, for instance, `https://climateatlas.ca/variables`). It is computed using the cumulative sum of the precipitation corrected variable (PRECTOTCORR, expressed as $kg/m^2/s$ and converted to $mm/day$) over each three consecutive days period. Our first indicator corresponds to the maximum value observed over a period of one year at the cell level.

**Climate warning:** High, persistent temperatures increase the risk of drought, which can severely impact food production and increase the risk of wildfire. High temperatures can also lead to more thunderstorms, which means increased risks of flash flooding, lightning, hail, and perhaps even tornadoes. We consider, as a measure of high temperature, the annual number of days satisfying conventional heatwave criteria. The World Meteorological Organization defines a heatwave as five or more consecutive days during which the daily maximum temperature (variable T2MMAX in Kelvin, converted to Celsius) exceeds the average maximum temperature by $5°C$ or more. The average maximum temperatures were calculated using a running window of $+/-$ 7 days centered on each day of the year for the benchmark climatology decade from 1980 to 1989. Our second indicator corresponds to the total annual count of days satisfying the heatwave criteria at the cell level.

Figure 1 displays the boxplot of the two indicators computed for each year on the $207,936$ cells, along with the yearly quantile plots at levels 0.5, 0.95, 0.99, 0.995, and 0.999. While it is difficult to observe a substantial change in the interquartile range of the precipitations' distribution, the quantile values at hight levels tend to increase significantly across time. As regards heatwaves, although the entire distribution appears to be shifting toward higher values, it is the large-order quantiles that experience the largest increases. This motivates the need for using statistical tools to measure the evolution of extreme values in the data.

# 3 From quantile to extremile risk assessment

Daouia et al. (2019, 2022) have introduced an alternative class to quantiles, called extremiles, which suggest better capability of fitting spread in data points and displaying interesting features of heavy-tailed distributions. Let $Y$ be a random variable with continuous distribution function $F$. Given a $\tau \in (0,1)$, the **quantile** $q_\tau$ of $F$ can uniquely be defined through

$$\begin{aligned} F^{-1}(\tau) &= \inf\{y : F(y) \geq \tau\} \\ &\equiv \text{median}(Z_\tau), \end{aligned}$$

where the random variable $\boldsymbol{Z_\tau}$ has c.d.f. $F_{Z_\tau} = K_\tau(F)$ with $K_\tau(t) = t^{r(\tau)}$ and $r(\tau) = \frac{\log(1/2)}{\log(\tau)}$. The **extremile** of order $\tau$ of $Y$ is then defined as the expectation of $Z_\tau$, $\mathbb{E}(Z_\tau)$, that is,

$$\xi_\tau = \mathbb{E}\left[Y\, J_\tau(F(Y))\right] = \int_0^1 J_\tau(t)\, q_t\, dt = \int_0^1 q_t\, dK_\tau(t), \tag{1}$$

where $J_\tau(t) = K'_\tau(t) = r(\tau)\, t^{r(\tau)-1}$. When $r(\tau)$ is a positive integer, we then have

$$\xi_\tau = \mathbb{E}\left[\max(Y_1, \ldots, Y_{r(\tau)})\right], \tag{2}$$

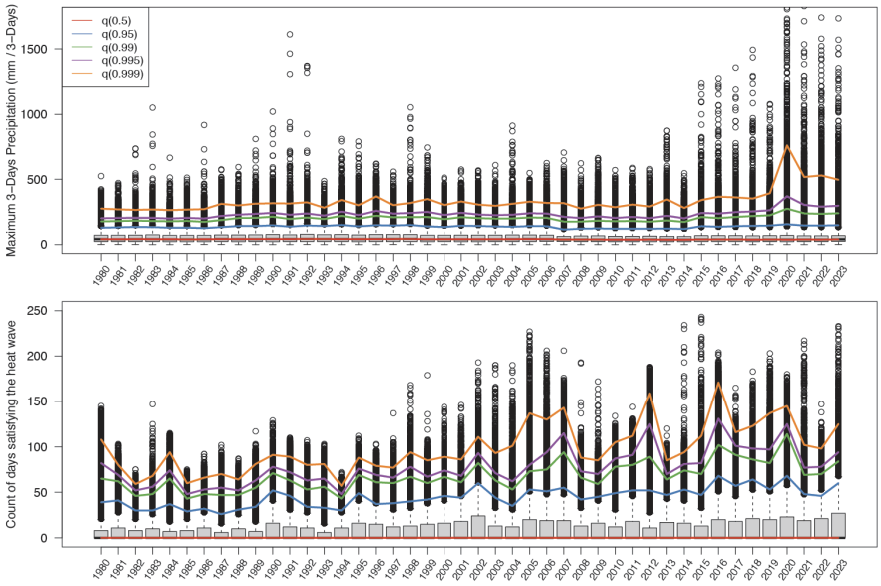



Figure 1: Boxplot of the two indicators (maximum consecutive three-day precipitation on the top and number of days satisfying the heatwave criteria on the bottom) computed for each year on the 207, 936 cells. The quantile plots (of order 0.5, 0.95, 0.99, 0.995 and 0.999) indicate an increase in extreme values.

for independent copies $Y_i$ of $Y$. In our assessments with annual data, the extremile of order $\tau = 0.95$ has the intuitive meaning as the average maximum value that we would obtain every 14 years.

Extremiles also define a least squares analog

$$\xi_\tau = \arg\min_\theta \mathbb{E}\left\{J_\tau(F(Y)) \cdot |Y - \theta|^2\right\}$$

to quantiles

$$q_\tau = \arg\min_\theta \mathbb{E}\left\{J_\tau(F(Y)) \cdot |Y - \theta|\right\},$$

by substituting the squared deviations in place of the absolute deviations. As such, they are motivated via several angles, which reveals their specific merits and strengths, including the following properties:

- The existence of extremiles requires a finite first moment.
- Extremiles have a straightforward interpretation in terms of expected maxima.
- Extremiles are law–invariant and have an explicit integral representation in terms of the quantile function ($L$-statistic).

- Extremiles take into account the whole of the distribution: they depend on the value of each observation and give more weight to extremes (while quantiles only depend on the relative frequency of tail observations).
- Extremiles define a coherent and comonotonically additive risk measure, which is more conservative (pessimistic) than the quantile-Value at Risk (VaR).

# 4 Results

## 4.1 At the cell level

We have 44 observations (one observation per year) at each of the 207,936 cells. The top panels of Figures 2 and 3 show the cartography of the sample extremiles computed at level $\tau = 0.95$ for our two indicators. We indicate in small triangles the 100 zones with the highest tail exposure. For the top ten risky zones with the highest estimated extremiles, the bottom panels of Figures 2 and 3 exhibit the evolution of both sample $\tau$th extremiles and quantiles for $\tau \geq 0.80$.

One may wonder wether extremiles are high because the indicators are usually abundant (high values every year) or if this is only due to an exceptional value in a given year. It is exactly for this reason why we superimpose the quantile and extremile plots, for the top 10 risky cells, in the bottom panels of Figures 2 and 3. Another interesting question is to track how do risk measures change over time. Since we only have 44 annual observations per cell, we will answer this question below, in subsection 4.2, with data aggregated at the country level so as to have larger datasets.

**Heavy precipitation:** We can see in Figure 2 (top) that the risk areas for heavy precipitation are mainly located between latitude $-30°$ and latitude $30°$, which unsurprisingly corresponds to areas with a tropical climate characterized by heavy precipitation in the summer. Among the Top 10 values, we can observe in Figure 2 (bottom) that some locations are risky because there are heavy precipitations every year (Top 1, 3, 4, or 5 that correspond to cells located in the Independent State of Papua New Guinea, Kenya, Nigeria, and India). By contrast, we observe that some zones (Top 2, 6, or 7 that correspond to cells located in Mexico, Morocco, and Angola) are risky because of only one or two extreme events that influence the extremiles. It is interesting to see that in such heavy-tailed cases, sample quantiles remain much less alert to these catastrophic events than their extremile analogs.

**Climate warning:** We can see in Figure 3 (top) that the most impacted zones by climate warning are localized near the poles, in South America (mostly in Venezuela and Brazil), in Central Africa (the Democratic Republic of the Congo, Angola), continental Europe, and North America. It may seem surprising that certain regions or countries, such as the Sahara or India, do not appear among the risky regions. We recall that heatwaves occur when temperatures exceed the normal thresholds observed between 1980 and 1989 by $5°C$. As

temperatures were already high between 1980 and 1989 in these areas, even if temperatures have increased by a few degrees in recent years, they did not exceed $5°C$. Among the Top 10 values, we can observe in Figure 3 (bottom) that all locations are risky not because of just one or two rare events, but due to frequent heatwaves in these zones, which explains why tail quantiles and extremiles are very close (it is established in Daouia et al., 2019, that tail quantiles and extremiles are equivalent for light-tailed distributions).

### 4.2 At the country level

We aim to compute risk measures at the country level and see their evolution over time. In doing so, we consider all cells falling within the borderlines of a given country. A country contains, on average, 240 cells. Russia is the country with the largest number, with more than 10,000 cells. Note that each country has at least 10 cells (for small countries, such as island countries, we add the nearest neighbors until we have 10 cells by country).

Figures 4 and 5 showcase the sample extremiles computed at the country level for four distinct 10-year periods: 1984-1993, 1994-2003, 2004-2013, and 2014-2023. As visualized from Figure 4, precipitations tend to increase across time in the countries close to the tropics. This is particularly visible for Kenya in Central Africa, Bangladesh in South Asia, Nicaragua in Central America, or the Philippines in Southeast Asia, where the highest extremile measurements are observed in 2023. In Bangladesh, the risk of heavy rainfall over 3 days has increased from around $200mm$ thirty years ago to $700mm$ nowadays.

We remark from Figure 5 that extremiles of heatwaves tend to increase across time everywhere, in particular near the poles, in Central Africa (with a peak observed during the 2004-2013 period), northern South America, Europe and North America. For example, in Venezuela, the risk of heatwave has increased from around 5 days per year thirty years ago to 135 days nowadays.

## Bibliographie

Daouia A., Gijbels I. and Stupfler G. (2019), Extremiles: A New Perspective on Asymmetric Least Squares, *Journal of the American Statistical Association*, 114(527), pp. 1366–1381.

Daouia A., Gijbels I. and Stupfler G. (2022), Extremile Regression, *Journal of the American Statistical Association*, 117(539), pp. 1579–1586.

Fischer, E. and Knutti, R. (2016), Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, 6, pp. 986—991.

Gelaro R., McCarty W., Suárez M.J., Todling R., Molod A., Takacs L., Randles C.A., Darmenov A., Bosilovich M. G., Reichle R., Wargan K., Coy L., Cullather R., Draper C., Akella S., Buchard V., Conaty A., da Silva A. M., Gu W., Kim G.-K., Koster R., Lucchesi R., Merkova D., Nielsen J. E., Partyka G., Pawson S., Putman W., Rienecker M., Schubert S.D., Sienkiewicz M., and Zhao B. (2017) The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of Climate*, 30(14), pp. 5419–5454.

Katz R.W. and Brown B.G. (1992), Extreme events in a changing climate: variability is more important than averages. *Climatic change*, 21(3), pp. 289–302.

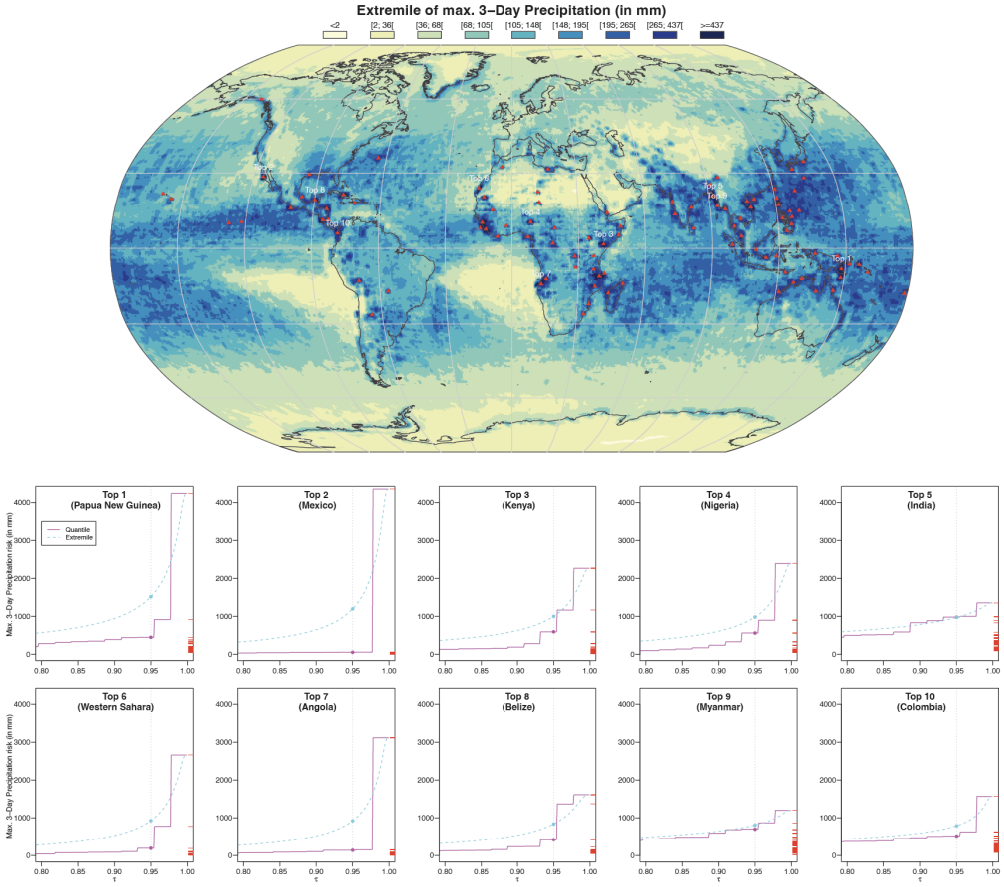



Figure 2: (Top) Map of the sample 0.95th extremile computed at the cell level for the heavy precipitation. The small triangles in red represent the 100 areas with the highest risk values. (Bottom) Plots of the quantile and extremile risks as functions of $\tau \in [0.80, 1]$ for the top ten risky zones with the highest estimated extremiles.

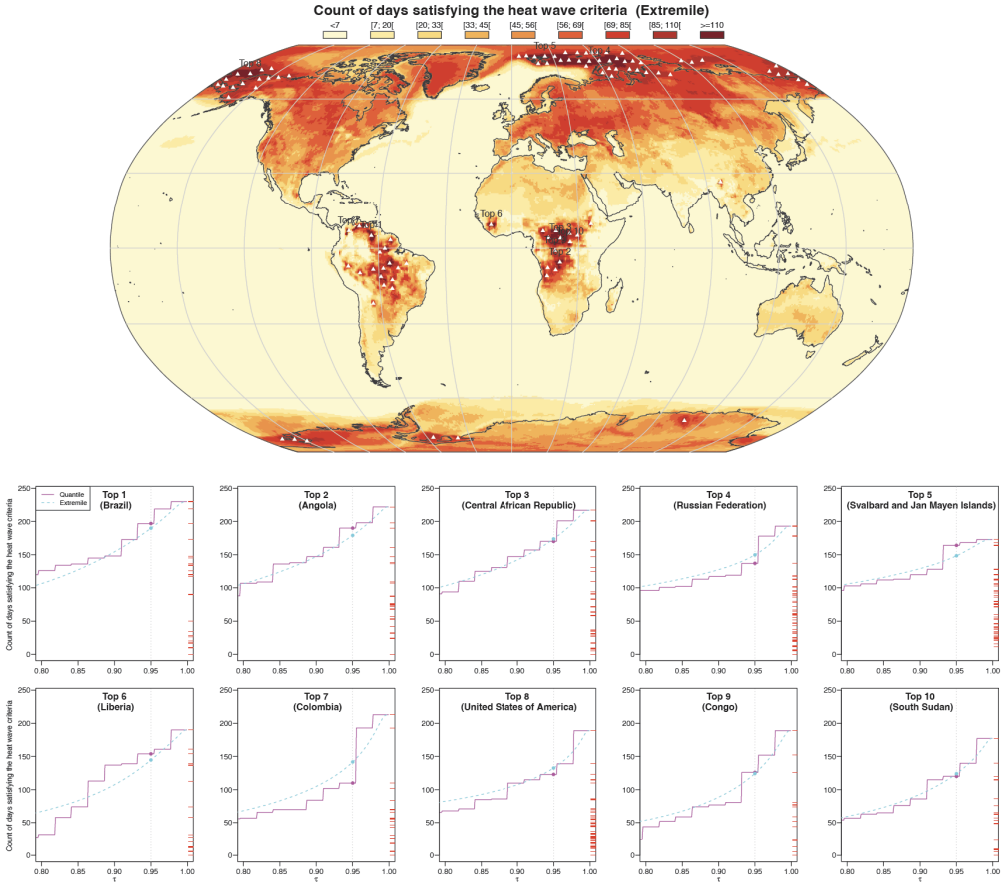



Figure 3: (Top) Map of the sample 0.95th extremiles computed at the cell level for the climate warning indicator. The small triangles in white represent the 100 areas with the highest risk values. (Bottom) Plots of the quantile and extremile risks as functions of $\tau \in [0.80, 1]$ for the top ten risky zones with the highest estimated extremiles .

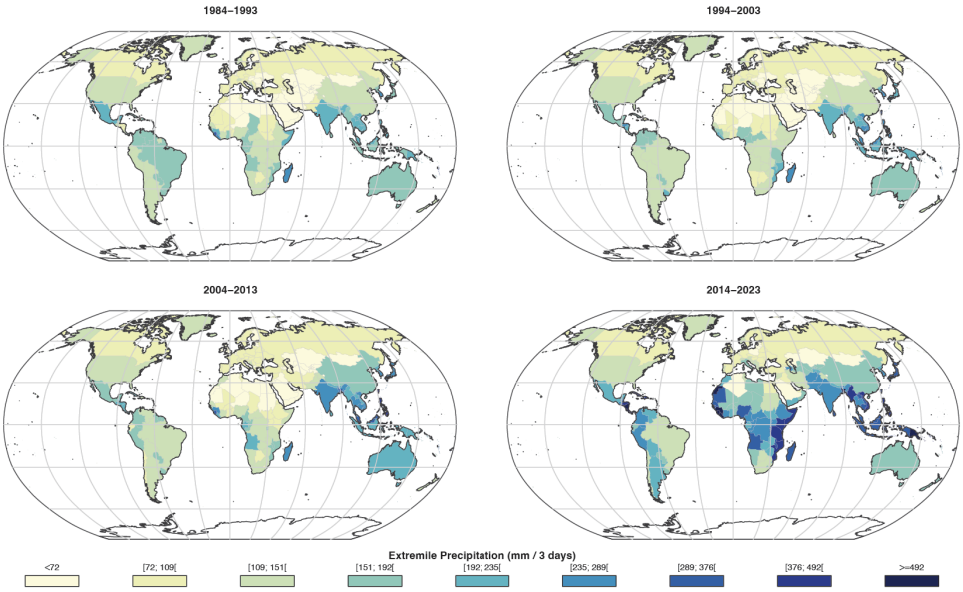



Figure 4: Estimated extremiles for heavy precipitations by country and over periods of 10 years

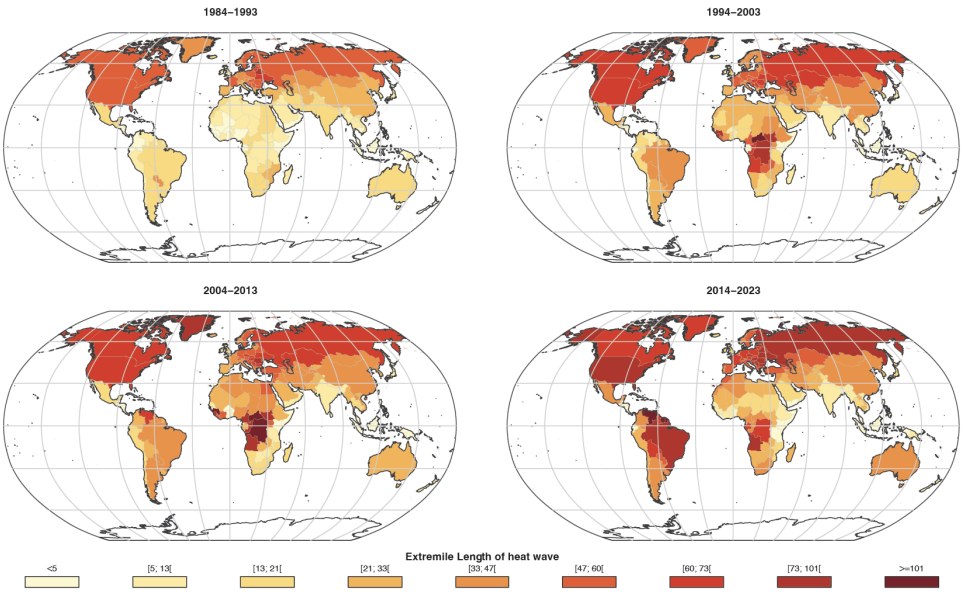



Figure 5: Estimated extremiles for heatwaves by country and over periods of 10 years

# Transport optimal

# Estimation d'une distance de Wasserstein par transport optimal entropique

Jérémie Bigot [1], Paul Freulon [2], Boris P.Hejblum [3] & Arthur Leclaire [4]

[1] *Université de Bordeaux, Bordeaux, 33000, France. Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.*
*jeremie.bigot@math.u-bordeaux.fr*

[2] *EPFL, Institut de Mathématiques, CH-1015 Lausanne, Suisse. Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.*
*paul.freulon@epfl.ch*

[3] *Université de Bordeaux, Bordeaux, 33000, France. Bordeaux Population Health Research Center Inserm U1219, Inria SISTM, 33000 Bordeaux, France. Vaccine Research Institute (VRI), 94010 Créteil, France.*
*boris.hejblum@u-bordeaux.fr*

[4] *LTCI, Télécom Paris, IP Paris, 91120 Palaiseau, France. Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.*
*arthur.leclaire@telecom-paris.fr*

**Résumé.** Le transport optimal entropique a été introduit en 2013 par M.Cuturi pour accélérer le calcul des distances de transport optimal. Dans cette présentation, on cherche à estimer une distance de Wasserstein en utilisant ce transport optimal régularisé. L'estimateur étudié est une distance de transport entropique où les deux mesures ont été remplacées par leurs versions empiriques. On présente des résultats de convergence de cet estimateur vers la distance Wasserstein évaluée entre les mesures sous-jacentes aux observations. Dans un premier temps, on essaye de motiver ce problème par une question classique de statistique : comment ajuster un modèle statistique à des observations ? Dans une deuxième partie, on fait l'état de l'art de résultats permettant de contrôler l'erreur de l'estimateur considéré. Enfin, on présente un nouvelle borne d'erreur pour l'estimateur étudié. De cette borne, on déduit un choix du paramètre de régularisation.

**Mots-clés.** Transport optimal, régularisation entropique, vitesse de convergence.

**Abstract.** For a faster computation of the optimal transport problem, M.Cuturi introduced in 2013 the entropic optimal transport. In this communication, we estimate a Wasserstein distance thanks to this regularized optimal transport. More precisely, we substitute the compared measures by their empirical counterparts in the regularized Wasserstein distance. We present rates of convergence of this estimator towards the Wasserstein distance between the underlying measures. Firstly, we try to motivate this estimation problem with a standard statistical question: how to fit a statistical model to some observations? Secondly, we review some known results that enable us to control the estimation error. Finally, we present a new bound on the estimation error. From this new bound, we deduce a choice for the regularization parameter.

**Keywords.** Optimal transport, entropic regularization, rate of convergence.

# 1 Notations, motivations, et problème étudié

Dans ce texte, on travaille dans l'espace euclidien à $d$ dimensions $\mathbb{R}^d$. Soit $Y_1, \dots, Y_n \in \mathbb{R}^d$ une série de $n$ observations supposées indépendantes et identiquement distribuées. Un problème classique est d'approcher la loi de $Y_1, \dots, Y_n$ par une collection de mesures $\mathcal{P} := \{\mu_\theta \ : \ \theta \in \Theta\}$. En termes statistiques, on parle d'un modèle $\mathcal{P}$ paramétré par l'ensemble $\Theta$. Une approche basée sur la fonction de vraisemblance permet d'aborder ce problème. Sous certaines hypothèses, la fonction de vraisemblance s'interprète comme une version empirique de la divergence de Kullback-Leibler [13, Thm. 9.13].

**Definition 1.1.** *Soit $\mu$ et $\nu$ deux mesures de probabilité sur $\mathbb{R}^d$. La divergence de Kullback-Leibler entre $\mu$ et $\nu$ est définie par*

$$\mathrm{KL}(\mu|\nu) = \int_{\mathbb{R}^d} \log\Big(\frac{d\mu}{d\nu}\Big) d\mu,$$

*si $\mu$ est absolument continue par rapport à $\nu$. Sinon, $\mathrm{KL}(\mu|\nu) = +\infty$.*

Les approches reposant sur la fonction de vraisemblance, ou la divergence de Kullback-Leibler, nécessitent des hypothèses d'absolue continuité. Retirer ces hypothèses motive l'utilisation des distances de transport en statistique [1, 7].

**Definition 1.2.** *Soit $\mu$ et $\nu$ deux mesures de probabilité sur $\mathbb{R}^d$ admettant des moments d'ordre deux. En notant $\Pi(\mu, \nu)$ l'ensemble des mesures sur $\mathbb{R}^d \times \mathbb{R}^d$ ayant pour marginales $\mu$ et $\nu$, la 2-distance de Wasserstein $W_0(\mu, \nu)$ est définie par*

$$W_0(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x - y\|^2 d\pi(x, y), \tag{1.1}$$

*où $\|x - y\|$ est la distance euclidienne entre $x$ et $y$.*

La quantité $W_0$ introduite en équation (1.1) permet de définir une distance entre mesures de probabilité [12]. Le calcul numérique de la distance $W_0$ est relativement lent. C'est pour en accélérer le calcul que M.Cuturi propose de régulariser le problème d'optimisation (1.1). À notre connaissance, c'est dans l'article [4] qu'est introduit pour la première fois la distance de Wasserstein régularisée.

**Definition 1.3.** *Soit $\mu$ et $\nu$ deux mesures de probabilité sur $\mathbb{R}^d$; et $\lambda \geq 0$ un paramètre de régularisation. La distance de Wasserstein régularisée $W_\lambda$ est définie par le problème*

$$W_\lambda(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x - y\|^2 d\pi(x, y) + \lambda \,\mathrm{KL}(\pi|\mu \otimes \nu). \tag{1.2}$$

Remarquons que lorsque $\lambda = 0$, on retrouve la distance de Wasserstein classique (1.1). D'autres régularisations que la divergence de Kullback-Leibler ont été proposées [8]. Pour le distinguer des autres régularisations, le problème (1.2) est parfois désigné par transport optimal entropique. Enfin, même si on parle de distance lorsque $\lambda > 0$, le problème de

optimal régularisé ne définit plus une distance. En effet, sur la droite réel $\mathbb{R}$, en posant $\mu = (\delta_0 + \delta_1)/2$, on a $W_\lambda(\mu, \mu) > 0$ lorsque $\lambda > 0$.

Les critères d'écart entre lois de probabilité étant introduits, revenons à des considérations statistiques. On va se concentrer sur l'utilisation des distances de transport (1.1) et (1.2). Supposons nos observations $Y_1, \dots, Y_n$ identiquement distribuées selon une mesure inconnue $\nu$. En utilisant une distance de transport comme critère d'écart, approcher la mesure $\nu$ par le modèle $\{\mu_\theta \mid \theta \in \Theta\}$ nécessite de résoudre le problème $\min_{\theta\in\Theta} W_0(\mu_\theta, \nu)$.

Dans ce texte; plutôt que d'étudier le problème $\min_{\theta\in\Theta} W_0(\mu_\theta, \nu)$, on va aborder une question préliminaire: estimer $W_0(\mu, \nu)$. Comme nous sommes dans un problème statistique, la mesure $\nu$ est inconnue. Dans certains cas [6], la mesure $\mu$ n'est pas connue non plus. C'est ce qu'on va supposer dans la suite de ce texte. Soit $X_1, \dots, X_n \sim \mu$; et soit $Y_1, \dots, Y_n \sim \nu$. On va substituer $\mu$ et $\nu$ par leurs mesures empiriques respectivement définies par

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \quad \text{et} \quad \hat{\nu}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_i}.$$

De plus, afin d'accélérer le calcul, on va remplacer la distance $W_0$ par sa version régularisée $W_\lambda$. Ainsi, alors que l'on souhaite étudier la quantité $W_0(\mu, \nu)$, on va calculer $W_\lambda(\hat{\mu}_n, \hat{\nu}_n)$. Autrement dit, $W_\lambda(\hat{\mu}_n, \hat{\nu}_n)$ est un estimateur de la quantité d'intérêt $W_0(\mu, \nu)$. C'est pourquoi nous cherchons à contrôler l'écart

$$\mathbb{E}[|W_0(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|]. \tag{1.3}$$

Dans la suite de ce texte, nous allons discuter de cette question. En section 2 on examine certains résultats déjà établis permettant de contrôler (1.3). Puis, en section 3, on propose de nouveaux résultats permettant de majorer l'erreur (1.3). Finalement, on déduit de la majoration obtenue un choix de paramètre de régularisation. Tous nos résultats nécessitent de supposer que les mesures $\mu$ et $\nu$ ont des supports bornés. Plus précisément, on va supposer que les supports de $\mu$ et $\nu$ sont inclus dans la boule centrée de rayon $R$. On note cette boule $B(0, R) := \{x \in \mathbb{R}^d \ : \ \|x\| \leq R\}$. De plus, toutes les observations considérées sont supposées indépendantes.

# 2 État de l'art

Une première étape pour contrôler l'erreur (1.3) est de la décomposer entre un terme d'approximation et un terme d'estimation. On commence donc par l'inégalité

$$|W_0(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)| \leq \underbrace{|W_0(\mu, \nu) - W_\lambda(\mu, \nu)|}_{\text{Approximation}} + \underbrace{|W_\lambda(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|}_{\text{Estimation}}. \tag{2.1}$$

Pour l'erreur d'approximation, on exploite le résultat suivant.

**Theorem 2.1.** *[5, Thm. 1] Soit $\lambda > 0$. Si $\mu$ et $\nu$ ont leur supports inclus dans la boule centrée de rayon $R$, alors*

$$|W_0(\mu,\nu) - W_\lambda(\mu,\nu)| \leq 2d\lambda \log\left(\frac{8e^2R^2}{\sqrt{d}\lambda}\right). \tag{2.2}$$

Dans le même article est établi l'erreur d'estimation (désignée par "sample complexity" dans [5]) que l'on rappelle ci-dessous.

**Theorem 2.2.** *[5, Thm. 3] Soit $\lambda > 0$. Si $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$ sont respectivement distribuées selon $\mu$ et $\nu$, alors*

$$\mathbb{E}[|W_\lambda(\mu,\nu) - W_\lambda(\hat{\mu}_n,\hat{\nu}_n)|] \lesssim \left(1 + \frac{1}{\lambda^{\lfloor d/2 \rfloor}}\right)\frac{1}{\sqrt{n}}, \tag{2.3}$$

*le symbole $\lesssim$ cachant une constante multiplicative dépendant de $R$ et $d$.*

Remarquons que dans [5], un facteur $e^{R^2/\lambda}$ était présent dans la majoration (2.3). Ce facteur $e^{R^2/\lambda}$ a été enlevé dans les travaux ultérieurs [10, 3]. En utilisant la décomposition (2.1), ainsi que les théorèmes 2.1 et 2.2, si la dimension $d$ est paire, on obtient le contrôle

$$\mathbb{E}\left[|W_0(\mu,\nu) - W_\lambda(\hat{\mu}_n,\hat{\nu}_n)|\right] \lesssim \lambda \log(\lambda^{-1}) + \lambda^{-d/2} n^{-1/2}.$$

Si l'on essaye d'optimiser en $\lambda$ le membre de droite de cette dernière inégalité, on obtient $\mathbb{E}\left[|W_0(\mu,\nu) - W_\lambda(\hat{\mu}_n,\hat{\nu}_n)|\right] \lesssim n^{-1/(d+2)} \log(n)$. Cette vitesse de convergence est largement plus lente que celle obtenue avec l'estimateur *non* régularisé $W_0(\hat{\mu}_n,\hat{\nu}_n)$. La vitesse de convergence de cet estimateur a été établi par Chizat et al. en 2020. Dans un soucis de concision, on rappelle ce résultat uniquement dans le cas où la dimension $d$ est strictement supérieure à quatre.

**Theorem 2.3.** *[3, Thm. 2] Soit $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$ deux séries de $n$ observations respectivement distribuées selon $\mu$ et $\nu$. Si la dimension $d$ des observations est strictement supérieure à quatre, on a alors*

$$\mathbb{E}[|W_0(\mu,\nu) - W_0(\hat{\mu}_n,\hat{\nu}_n)|] \lesssim n^{-2/d}.$$

Cette vitesse de convergence décroît lorsque la dimension augmente. On peut donc s'interroger sur l'existence d'un estimateur convergeant plus rapidement. Sans hypothèses supplémentaires sur les mesures considérées, la réponse à cette question est négative.

**Theorem 2.4.** *[9, Thm. 22] Sous l'hypothèse que $n$ observations indépendantes sont disponibles pour chaque mesure $\mu$ et $\nu$, on a*

$$(n\log(n))^{-2/d} \lesssim \inf_{\widehat{W}_n} \sup_{\substack{\mu\in\mathcal{P}(\mathbb{R}^d),\\ \nu\in\mathcal{P}(\mathbb{R}^d)}} \mathbb{E}\left[\left|W_0(\mu,\nu) - \widehat{W}_n\right|\right], \tag{2.4}$$

*où la borne inférieure est calculée sur l'ensemble des estimateurs de $W_0(\mu,\nu)$.*

Le théorème 2.4 est un résultat de type "minimax". On en déduit que la vitesse de convergence en $n^{-2/d}$ de l'estimateur non régularisé $W_0(\hat{\mu}_n,\hat{\nu}_n)$ est, en considérant le pire scénario, optimal.

# 3 Nouveaux résultats

Cette section présente nos principaux résultats concernant la convergence de $W_{\lambda_n}(\hat{\mu}_n, \hat{\nu}_n)$ vers $W_0(\mu, \nu)$. On va montrer qu'un paramètre de régularisation adapté permet d'atteindre (à un facteur logarithmique près) la vitesse minimax $n^{-2/d}$. Pour un texte plus détaillé, la lecture de la prépublication [2] est proposée.

L'amélioration de la vitesse de convergence que l'on obtient repose sur un nouveau contrôle de l'erreur d'estimation $|W_\lambda(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|$. Ce contrôle ne dépend pas du paramètre de régularisation.

**Proposition 3.1.** *[2, Prop. 3.1] Soit $\lambda \geq 0$. Si $n$ observations indépendantes sont disponibles pour les mesures $\mu$ et $\nu$, dans le cas où la dimension $d$ des observations est strictement supérieure à quatre, on a*

$$\mathbb{E}[|W_\lambda(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|] \lesssim n^{-2/d}, \tag{3.1}$$

*où $\lesssim$ cache une constante multiplicative dépendant de $R$ et de la dimension $d$.*

En combinant ce nouveau contrôle de l'erreur d'estimation avec l'erreur d'approximation donnée par le théorème 2.1, on obtient l'inégalité

$$\mathbb{E}[|W_0(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|] \lesssim n^{-2/d} + \lambda \log(\lambda^{-1}). \tag{3.2}$$

Dans cette dernière inégalité (3.2), nous n'avons pas de contrôle sur l'erreur d'estimation $n^{-2/d}$. En revanche, l'erreur d'approximation de l'ordre $\lambda \log(\lambda^{-1})$ est contrôlée par le paramètre de régularisation. On va donc choisir ce paramètre $\lambda$ de façon à obtenir une erreur d'approximation de même ordre de grandeur que l'erreur d'estimation.

**Theorem 3.1.** *[2, Prop. 4.2] Si $n$ observations indépendantes sont disponibles pour chacune des deux mesures $\mu$ et $\nu$, alors*

$$\mathbb{E}[|W_0(\mu, \nu) - W_{\lambda_n}(\hat{\mu}_n, \hat{\nu}_n)|] \lesssim n^{-2/d} \log(n) \quad \textit{avec} \quad \lambda_n = n^{-2/d},$$

*où $\lesssim$ cache une constante multiplicative dépendant uniquement de $R$ et $d$.*

Ce dernier résultat montre qu'un choix de paramètre $\lambda_n$ dépendant du nombre d'observations disponibles permet atteindre, à un facteur logarithmique près, la vitesse minimax avec l'estimateur $W_{\lambda_n}(\hat{\mu}_n, \hat{\nu}_n)$. On pourrait choisir un paramètre de régularisation encore plus petit que $n^{-2/d}$; mais un tel choix ralentirait le calcul numérique de $W_\lambda(\hat{\mu}_n, \hat{\nu}_n)$. L'algorithme de Sinkhorn [11] permet le calcul de coût de transport régularisé $W_\lambda$ dans le cas où les mesures comparées sont discrètes. La vitesse de convergence de cet algorithme ralentit lorsque le paramètre de régularisation décroit.

# References


[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[2] J. Bigot, P. Freulon, B. P. Hejblum, and A. Leclaire. On the potential benefits of entropic regularization for smoothing wasserstein estimators. *arXiv preprint arXiv:2210.06934*, 2022.

[3] L. Chizat, P. Roussillon, F. Léger, F. Vialard, and G. Peyré. Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. In *Proc. NeurIPS'20*, 2020.

[4] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[5] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[7] M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15(1):1328 – 1371, 2021.

[8] D. A. Lorenz, P. Manns, and C. Meyer. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3):1919–1949, 2021.

[9] T. Manole and J. Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108–1135, 2024.

[10] G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[12] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[13] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2004.

# Exploring Optimal Transport in Jazz Music Analysis Application to the Real Book

Jean Dufraiche [1] & Valérie Garès [2] & Madison Giacofci [3]& Nicolas Klutchnikoff [4]

[1] *INSA Rennes, France; jean.dufraiche@insa-rennes.fr*
[2] *Univ Rennes, INSA, IRMAR - UMR 6625, France; valerie.gares@insa-rennes.fr*
[3] *Univ Rennes, IRMAR - UMR 6625, France; joyce.giacofci@univ-rennes2.fr*
[4] *Univ Rennes, IRMAR - UMR 6625, France; nicolas.klutchnikoff@univ-rennes2.fr*

**Résumé.** Cet article se penche sur l'analyse mathématique du Real Book, un célèbre corpus de musique de jazz. Pour simplifier le problème, nous supposons que chaque pièce musicale est exprimée comme une séquence d'accords. Notre approche introduit une représentation des accords basée sur leurs emprunts aux différents modes pythagoriciens, une dimension qui semble négligée dans la littérature existante. Cette représentation innovante permet d'établir des dissimilarités entre les accords, ce qui constitue la base de la comparaison des pièces. Plus précisément, deux pièces de Real Book peuvent être représentées par la distribution empirique de leurs accords. La dissimilarité entre ces pièces est alors déterminée par le coût de transport optimal entre leurs distributions respectives. Ce calcul repose sur le coût entre accords défini précédemment.

**Mots-clés.** Analyse musicale par ordinateur, transport optimal, classification non supervisée

**Abstract.** This article delves into the mathematical analysis of the Real Book, a renowned collection of jazz music. To simplify the problem, we assume that each musical piece is expressed as a sequence of chords. Our approach introduces a chord representation based on their borrowings from various Pythagorean modes, a dimension that seems to be neglected in the existing literature. This innovative representation allows us to establish dissimilarities between chords, which form the basis for comparing songs. Specifically, two Real Book pieces can be represented by the empirical distribution of their chords, with dissimilarity determined as the optimal transport cost between them. This calculation is based on the previously defined chord costs.

**Keywords.** Music Information Retrieval, optimal transport, clustering

## 1 Introduction

When studying Western music with the use of symbolic music data (score, Midi, MusicXML, etc.), a central aspect that one would like to describe is the harmony. For this reason, numerous representations and distances of chords have been defined in the field of music information retrieval (see Tymoczko, 2006; Rocher, Robine, and Hanna, 2010, and references therein).

These modelings have been used for various applications, such as recommandation algorithms (Aucouturier and Pachet, 2002) or to improve the performance of chord label recognition algorithms on signal-based data (Mauch, Noland, and Dixon, 2009), and it is assessed that it could help define goodness-of-fit measures when producing chord recognition on audio signal (Oudre, Grenier, and Févotte, 2011).

The link between chords and modes is usually not considered when we define such geometric representations; therefore, in our work, we propose a representation and a distance that depend on the potential belonging of a chord to a family of modes. This representation and distance might be of great interest for machine learning problems for which we need a distance or dissimilarity on the chords being the explanatory variables. We could think about clustering in unsupervised situations or $k$-nearest neighbors in supervised situations.

We applied this representation on a chord progressions dataset (de Berardinis, Meroño-Peñuela, Poltronieri, and Presutti, 2023) to find similarity in music extracts.

The outline is the following: having a corpus of songs for which we know their sequences of chords, we compute empirical distributions of chords for each of these songs. Then, using the Wasserstein metric on these empirical distributions based on a cost that we define on chords as in Givens and Shortt (1984), we obtain a pairwise distance between two songs.

# 2 Representation of chords and modes

**Pitch and note.** We rely on the MTS (Midi) format. In this standard, each frequency $\nu$ is mapped to a real number, called the **pitch**, according to the following function:

$$p(\nu) := 69 + 12\log_2(\nu/440), \quad \forall \nu > 0.$$

Although $p$ is a real-valued function, it is constructed to take integer values for the pitches used in the theory of Western music. In the following, we therefore assume that any pitch belongs to $\mathbb{N}$.

If two frequencies $\nu_1$ and $\nu_2$ are such that $\log_2(\nu_1/\nu_2) \in \mathbb{Z}$, then $p(\nu_1) \equiv p(\nu_2) \mod 12$. This corresponds to the situation where the two sounds are highly consonant, because they are separated by several octaves. The **pitch class** of $p$, denoted by $[p]$ or simply $p$ if the context allows, is then defined as its class in $\mathbb{Z}_{12} \cong [\![0, 11]\!]$. In this paper, we use the term **note** as a synonym for pitch class, although it is generally used to designate a set of properties such as pitch and duration associated with performance elements. However, there is no ambiguity in our context. The following table shows the correspondence between notes in $\mathbb{Z}_{12}$ and their names in English and French.

| Note | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| English | C | C♯/D♭ | D | D♯/E♭ | E | F | F♯/G♭ | G | G♯/A♭ | A | A♯/B♭ | B |
| French | Do | Do♯/Ré♭ | Ré | Ré♯/Mi♭ | Mi | Fa | Fa♯/Sol♭ | Sol | Sol♯/La♭ | La | La♯/Si♭ | Si |

**Chord.** When multiple sounds are played simultaneously (e.g. several keys on the piano), the corresponding musical element is called a chord that is characterized, in this work, by the set of the notes it contains. In harmony analysis, though, a specific note of a chord, called the **root**, acts as a foundational anchor, providing a reference point for the other notes within the chord. This root serves as a musical cornerstone that influences the perception and interpretation of the entire chordal structure. The significance of the root lies in its ability to establish a sense of stability and grounding. In the remainder of this paper, we make the distinction between an **unrooted chord**, which is simply a set of notes, and the more subtle notion of **chord**, which includes the notion of root.

More precisely, an unrooted chord is a collection $\mathcal{C} \subset \mathbb{Z}_{12}$ that is identified with its *one-hot encoding*, namely, the binary vector $b = (b_0, \ldots, b_{11})$ defined by $b_i = 1$ if $i \in \mathcal{C}$ and $b_i = 0$ otherwise (see Fujishima, 1999). A rooted chord, or more simply a chord, is a pair consisting of a set $\mathcal{C} \subset \mathbb{Z}_{12}$ and a particular note $r \in \mathcal{C}$, called the root.

For example, the chords *Cmajor* and *Aminor* respectively have roots $C$ and $A$ and the terms *major* and *minor* entirely define the other notes contained in *Cmajor* and *Aminor*. We then use a similar notation to Harte, Sandler, Abdallah, and Gómez (2005) that uses two features: the **root** $r \in \mathbb{Z}_{12}$ and the **kind** $k \in \mathcal{K} := \{0,1\}^{11}$, defined below.

For given unrooted chord $\mathcal{C}$ and root $r \in \mathcal{C}$, let us describe how the **kind** $k$ is deduced from the one-hot encoding $b = (b_0, \ldots, b_{11})$. Let us consider the example of the unrooted chord $\mathcal{C} = \{0, 4, 7, 9\}$ that has the following one-hot encoding:

$$b = (\underbrace{1}_{C}, 0, 0, 0, \underbrace{1}_{E}, 0, 0, \underbrace{1}_{G}, 0, \underbrace{1}_{A}, 0, 0).$$

If the root of this chord is chosen to be $C$ ($r = 0$), then the other notes of $\mathcal{C}$ can be deduced from the positions of the non-zero elements in $b$, see equation above. Now, if the root of the chord is different, for example $A$ ($r = 9$), then by applying a circular shift on the vector $b$, we can obtain a new vector such that we can deduce the other notes using the same process:

$$(\underbrace{1}_{C}, 0, 0, 0, \underbrace{1}_{E}, 0, 0, \underbrace{1}_{G}, 0, \underbrace{\mathbf{1}}_{\mathbf{A}}, \mathbf{0}, \mathbf{0}) \mapsto (\underbrace{\mathbf{1}}_{\mathbf{A}}, \mathbf{0}, \mathbf{0}, \underbrace{1}_{C}, 0, 0, 0, \underbrace{1}_{E}, 0, 0, \underbrace{1}_{G}, 0)$$

The vector $b$ is shifted so that $b_0$ corresponds to the root of the chord. It allows one to determine the other notes of $\mathcal{C}$ with the computations $[0] = [9+3]$, $[4] = [9+7]$, $[7] = [9+10]$.

Notice that we did the same process twice, but it was trivial when $r = 0$. Then, we define the kind $k$ as the 11 last components of the circular shifted vector that put the root on the first component.

Let $b \in \{0,1\}^{12}$ be the one-hot encoding of a chord, and let $r \in \mathbb{Z}_{12}$ be the root of $b$. We denote the kind:

$$k = (k_1, \ldots, k_{11}) \in \mathcal{K},$$

where $k_i = 1$ if and only if there exists a note $p$ in the chord $\mathcal{C}$ such that $[r + i] = [p]$. Then a chord can be defined by the elements $(r, b)$ or $(r, k)$. In the following, we choose the second solution and define a chord as a pair $(r, k)$ consisting of a fundamental note $r \in \mathbb{Z}_{12}$ and a kind $k \in \mathcal{K}$.

**Mode.** When a musical segment uses only a subset of notes, say $\mathcal{P}$, in its compositional choices, whether it contains chords (simultaneous) or melodies (consecutive), the elements of $\mathcal{P}$ are often perceived in relation to a fundamental note. This provides a representation similar to that of chords, in the form of a pair $(r, \ell) \in \mathbb{Z}_{12} \times \mathcal{K}$. In this context, we refer to $\ell$ as a **mode** and to $(r, \ell)$ as a **rooted mode**. By definition, $\ell_i = 1$ if and only if there exists a note $p \in \mathcal{P}$ such that $r + i \equiv p \mod 12$.

The two most famous modes are undoubtedly *major* and *minor* which correspond to $\ell = (0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1)$ and $\ell = (0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0)$, respectively. We can also mention Debussy's unitonal mode, which corresponds to $\ell = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$, as well as Messiaen's limited transposition modes.

At this point, it should be noted that musicians do not necessarily attach the same importance to all the pitches of a mode. To take this property into account, we can introduce a weight $w = (w_1, \ldots, w_{11})$ to reflect the importance of the different pitches present in a mode. Although the choice of a weight may seem subjective, it is possible to use the style of a musical corpus to give a reasonable definition in such a context.

**Pythagorean modes.** Pythagorean modes are undoubtedly the seven most widely used modes in Western music. They can be visually represented using the white keys on a piano. For example, constructing the Dorian mode involves playing a scale starting from D and corresponds to $\ell = (0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0)$. In jazz, musicians often employ modal borrowings as a technique to infuse color and diversity into their performances. Pythagorean modes serve as a structured foundation for generating tension and resolution, allowing musicians to create complex harmonic progressions.

In this context, it is well-known that specific pitches hold greater significance than others. This applies to the tonic triad—composed of the third and fifth notes of the mode—and the two notes within the mode separated by a tritone, spanning 6 semitones. With this in mind, we propose the following weights associated with each Pythagorean mode.

| Name | $\ell$ | $10 \cdot w$ |
|---|---|---|
| Lydian | $(0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1)$ | $(0, 1, 0, 2, 0, 3, 2, 0, 1, 0, 1)$ |
| Ionian | $(0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1)$ | $(0, 1, 0, 2, 2, 0, 2, 0, 1, 0, 2)$ |
| Mixolydian | $(0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0)$ | $(0, 1, 0, 3, 1, 0, 2, 0, 1, 2, 0)$ |
| Dorian | $(0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0)$ | $(0, 1, 3, 0, 1, 0, 2, 0, 2, 1, 0)$ |
| Aeolian | $(0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0)$ | $(0, 2, 2, 0, 1, 0, 2, 2, 0, 1, 0)$ |
| Phrygian | $(1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0)$ | $(2, 0, 2, 0, 1, 0, 3, 1, 0, 1, 0)$ |
| Locrian | $(1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0)$ | $(1, 0, 2, 0, 1, 4, 0, 1, 0, 1, 0)$ |

Elaborating on the notion of modes and their associated weights, we define a **musical system** as the couple $(\mathbb{M}, \mathbb{W})$ consisting of both a family of $M$ modes $\mathbb{M} = \{\ell^1, \ldots, \ell^M\}$ and a family of $M$ weights $\mathbb{W} = \{w^1, \ldots, w^M\}$. Musical system will allow us, in the next section, to represent all the chords and define a distance between them.

# 3 Cost between chords

A strong link exists between chords and modes. For a given mode $\ell$, chords can be created by selecting both a root $r$ and a kind $k$ such that $k_i \leq \ell_i,\ \forall i \in [\![1, 11]\!]$. In contrast, for a given chord $(r, k)$, there are several possibilities of rooted modes $(r, \ell)$ from which this chord is extracted. The cost we define in the following takes full advantage of this link. To our knowledge, this is the first time modes have been used in this way.

**Representation of kinds in a musical system.** Let $\mathbb{M} = \{\ell^1, \ldots, \ell^M\}$ and $\mathbb{W} = \{w^1, \ldots, w^M\}$ be a musical system and let $k \in \mathcal{K}$ be a kind of chord. We define the representation of $k$ in this musical system as the vector $x_{\mathbb{M},\mathbb{W}}(k) \in \mathbb{R}^M$ whose $j$-th coordinate is the $w_j$-weighted $l_1$–norm of $k - \ell^j$, that is:

$$x_{\mathbb{M},\mathbb{W}}(k) = \begin{pmatrix} \|k - \ell^1\|_1 \\ \vdots \\ \|k - \ell^M\|_M \end{pmatrix} \quad \text{where} \quad \|k - \ell^j\|_j = \sum_{i=1}^{11} w_i^j |k_i - \ell_i^j|.$$

If the representation $x_{\mathbb{M},\mathbb{W}}(\cdot)$ is injective (which is the case for the Pythagorean musical system but usually depends on the choice of $\mathbb{M}$ and $\mathbb{W}$), then the function $d : \mathcal{K} \times \mathcal{K} \to \mathbb{R}^+$ defined by

$$d_{\mathbb{M},\mathbb{W}}(k, k') = \|x_{\mathbb{M},\mathbb{W}}(k) - x_{\mathbb{M},\mathbb{W}}(k')\| = \left( \sum_{j=1}^{M} \big( \|k - \ell^j\|_j - \|k' - \ell^j\|_j \big)^2 \right)^{1/2}$$

is a distance. Otherwise, $d$ is a pseudo-metric.

**Example in the Pythagorean musical system.** Figure 1 shows four examples of representations of very commonly used kinds. Without going to much into the music theory details, we know that the major kind is the tonic triad of the Ionian, Lydian and Mixolydian modes, the minor kind is the tonic triad of the Dorian, Phrygian, and Aeolian modes, and the diminished kind is the tonic triad of the Locrian mode. In these three cases, we can see that the representation does show smaller values for the corresponding modes. In particular, the mode of minimum cost always contains the represented kind. The dominant kind is a major triad with an additional note that is only in the Mixolydian mode, this diagram shows that the belonging of the dominant kind to the Mixolydian mode is more discriminant than it was with the major kind.

**Cost on chords with the same roots.** If two chords $(r, k)$ and $(r, k')$ share the same root, the cost between them is defined by

$$c_{\mathbb{M},\mathbb{W}}\big((r, k), (r, k')\big) = d_{\mathbb{M},\mathbb{W}}(k, k').$$

Figure 2 gives an example (in the Pythagorean musical system) of clustering using this cost, which motivates the fact that it shows substitutions of chord kinds that are very frequent in the jazz genre.

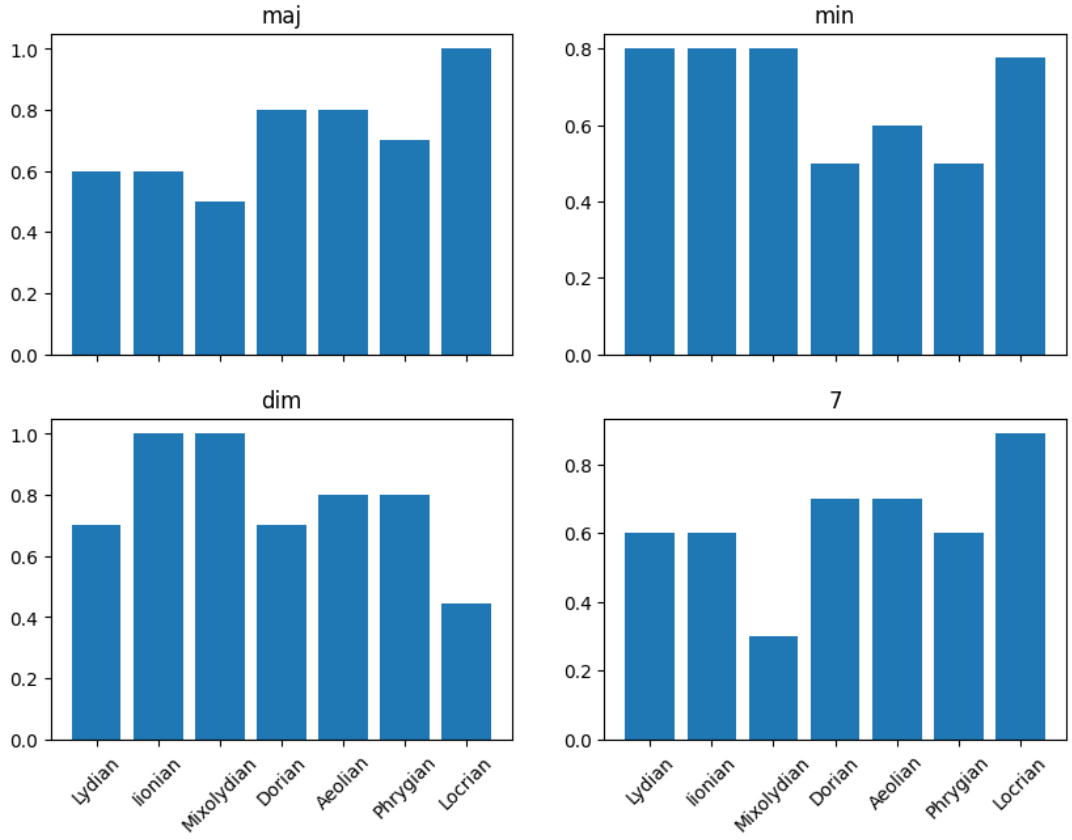



Figure 1: Representations of $x_{\mathbb{M},\mathbb{W}}(\cdot)$ for several kinds of chord: major, minor, diminished and dominant (i.e. 7) kinds.

**Cost on chords with different roots.** If the chords do not share the same roots, we consider that we should penalize the fact that playing one chord with the root of the other leads to a great difference between the original kind representation $x_{\mathbb{M},\mathbb{W}}(\cdot)$ and the new one. For example, the chords *Cmaj6* and *Amin7* have exactly the same notes $\{C, E, G, A\}$ but by choosing the roots $C$ and $A$ we obtain two different kinds, *maj6* and *min7*. The cost we propose will give a value of 0 for these two chords, as changing the root of *Cmaj6* to $A$ produces exactly *Amin7*.

To define this cost, let $(r^a, k^a)$ and $(r^b, k^b)$ be two chords. Denote by $\mathcal{C}^a$ the unrooted chord associated with $(r^a, k^a)$, that is, the notes that form this chord. We also consider the unrooted chord $\mathcal{C}^{ba} = \mathcal{C}^a \cup \{r^b\} \subset \mathbb{Z}_{12}$ and define $(r^b, k^{ba})$ as the chord $\mathcal{C}^{ba}$ with root $r^b$. The chord $(r^a, k^{ab})$ is defined in a similar way. These constructions allows us to define:

$$c_{\mathbb{M},\mathbb{W}}\big((r^a, k^a), (r^b, k^b)\big) = \Big[c_{\mathbb{M},\mathbb{W}}\big((r^a, k^a), (r^a, k^{ab})\big) + c_{\mathbb{M},\mathbb{W}}\big((r^b, k^b), (r^b, k^{ba})\big)\Big]/2.$$

Notice that $c_{\mathbb{M},\mathbb{W}}$ is a dissimilarity (which is not a distance in the general case) since for all

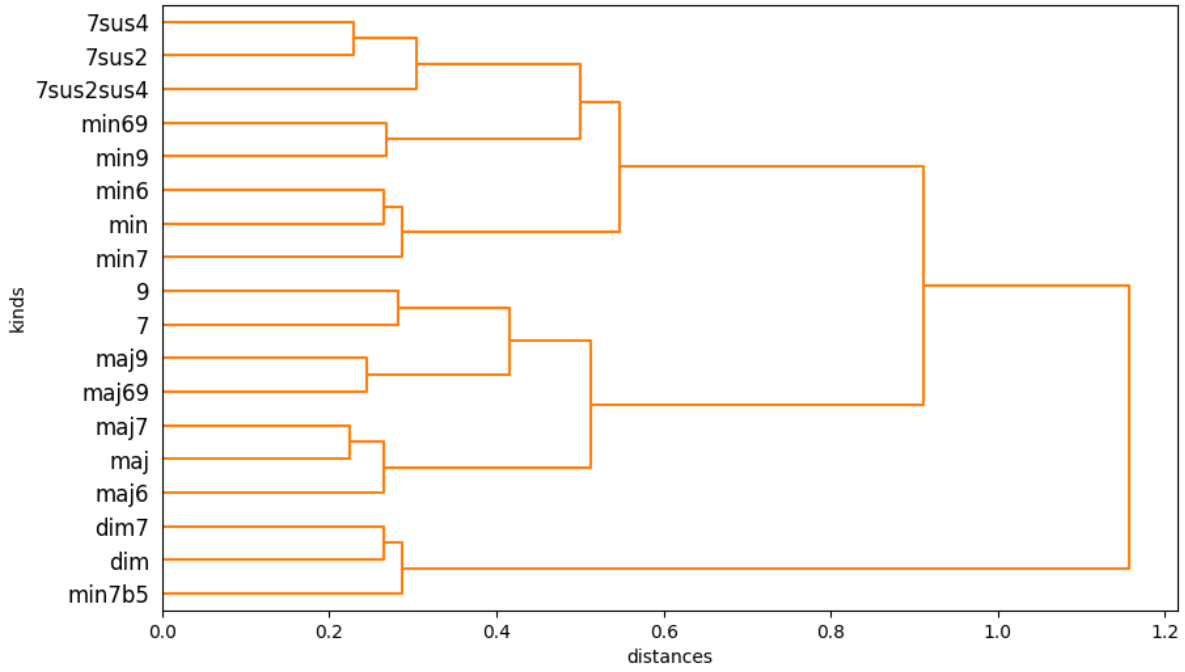



Figure 2: Hierarchical clustering performed on a small set of very often used kinds in the jazz repertoire

chords $(r^a, k^a)$ and $(r^b, k^b)$ in $\mathbb{Z}_{12} \times \mathcal{K}$ we have:

$$c_{\mathbb{M},\mathbb{W}}\big((r^a, k^a), (r^b, k^b)\big) = c_{\mathbb{M},\mathbb{W}}\big((r^b, k^b), (r^a, k^a)\big) \geq 0 \quad \text{and} \quad c_{\mathbb{M},\mathbb{W}}\left((r^a, k^a), (r^a, k^a)\right) = 0.$$

Figure 3 gives an example (in the Pythagorean musical system) of clustering that we produce using this dissimilarity on the chords that appear in the song *Cry Of The Wild Goose*[1]. This figure illustrates the fact that chords with a small dissimilarity are more likely to be substitutions of each other (adding $G$ to $B\flat major$ produces the chord $Gmin7$ and removing $F$ from $Gmin7$ produces $Gmin$).

# 4 Dissimilarity between songs

In our study, we assume that a song $S$ can be represented by a sequence of chords, that is, $S = \{(r^1, k^1), \ldots, (r^{n_S}, k^{n_S})\}$. As the number of chords can vary from one song to another, it is essential to find a representation that is independent of this number to compare two songs. We thus associate a probability measure, defined on $\mathbb{Z}_{12} \times \mathcal{K}$, to each song $S$ in the following

[1]The names of the chords are provided by the `commonName` function of the `music21` python package for computational Musicology (Cuthbert & Ariza, 2010).

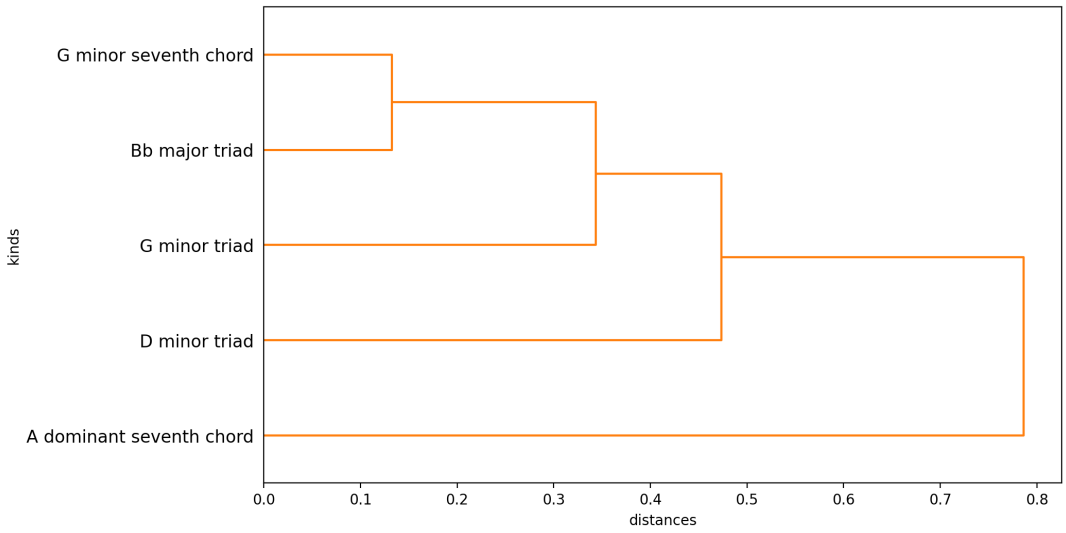



Figure 3: Hierarchical clustering performed on the chords of the song *Cry Of The Wild Goose*

way:

$$P_S(r,k) = \frac{1}{n_S}\sum_{i=1}^{n_S} 1_{[r^i=r,k^i=k]}, \quad \forall (r,k) \in \mathbb{Z}_{12}\times\mathcal{K}.$$

We then measure the proximity between two songs $S_1$ and $S_2$, as the optimal transport cost between $P_{S_1}$ and $P_{S_2}$, that is:

$$\delta_{\mathbb{M},\mathbb{W}}(S_1,S_2) = \left(\inf_{\gamma\in\Gamma(P_{S_1},P_{S_2})}\int_{E^2} c^2_{\mathbb{M},\mathbb{W}}(x,y)\,\mathrm{d}\gamma(x,y)\right)^{1/2},$$

where $E = \mathbb{Z}_{12}\times\mathcal{K}$ and $\Gamma(P_{S_1},P_{S_2})$ denotes the set of all joint probability measures $\gamma$ on $E^2$ whose marginals are $P_{S_1}$ and $P_{S_2}$. The Monge-Kantorovich problem (see Villani, 2009) mentioned above can be reformulated in a more explicit and equivalent manner as follows:

$$\delta^2_{\mathbb{M},\mathbb{W}}(S_1,S_2) = \min_{\gamma\in\Gamma(P_{S_1},P_{S_2})}\sum_{x\in E}\sum_{y\in E} c^2_{\mathbb{M},\mathbb{W}}(x,y)\gamma(x,y),$$

$$\text{subject to}\begin{cases}\displaystyle\sum_{y\in E}\gamma(x,y) = P_{S_1}(x), & \forall x\in E,\\ \displaystyle\sum_{x\in E}\gamma(x,y) = P_{S_2}(y), & \forall y\in E,\\ \gamma(x,y)\geq 0, & \forall (x,y)\in E^2.\end{cases}$$

Since $c_{\mathbb{M},\mathbb{W}}$ is a dissimilarity on chords, $\delta_{\mathbb{M},\mathbb{W}}$ is also a dissimilarity between songs. Thus, one can use specific data analysis tools, like hierarchical cluster analysis, which only require a dissimilarity matrix.

For a given song $S$, very few chords $(r,k)$ satisfy $P_S(r,k) > 0$ compared to the 24.576 elements of $E$. This sparsity property motivates the use of a dissimilarity that is defined using Optimal

Transport to take into account the fact that some chords can easily be transported to another one with a small cost. Indeed, many chords can be considered as substitutes one to the other (e.g., *Cmaj6* and *Amin7* that have a cost of 0).

# 5 Experiments

To produce our experiments and to test our dissimilarity between songs, we exclusively used the chord progression dataset compiled and published by de Berardinis et al. (2023). We limited our study to a total of 2.846 songs from the dataset, specifically those from the Real Book corpus (see Mauch, Dixon, Harte, Casey, and Fields, 2007).

Figure 4 shows the empirical distribution of the chords in the song *Cry Of The Wild Goose* from the Real Book (we can see the sparsity property discussed in the last section, as only five chords are used). We also represented two other songs, *Into It* and *Got A Match*, for which we computed the dissimilarity with respect to *Cry Of The Wild Goose*.

These results are easily interpreted from a harmonic point of view. *Got A Match* and *Cry Of The Wild Goose* share chords in common (or very close to each other) which leads to a smaller dissimilarity than between *Into It* and *Cry Of The Wild Goose* whose chords are not consonant, such as *D minor* and *D♭ minor 7*. This can be seen as a drawback of the definition of $\delta_{\mathbb{M},\mathbb{W}}$ that does not take into account possible transpositions. This point, as well as taking into account the harmonic evolution of the songs, is left for future research.

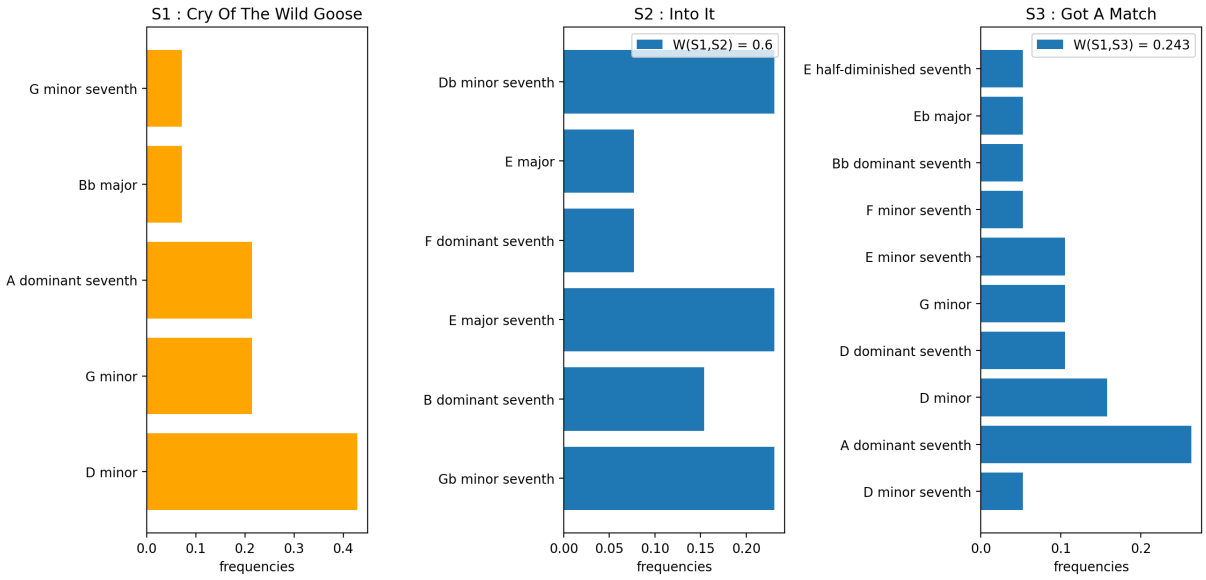



Figure 4: Histograms of the empirical distributions of chords in 3 songs of the Real Book

# References


Aucouturier, J.-J. and F. Pachet (2002). Music similarity measures: What's the use? In *International Society for Music Information Retrieval Conference*.

de Berardinis, J., A. Meroño-Peñuela, A. Poltronieri, and V. Presutti (2023). Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs. *Scientific Data 10*, 1–25.

Fujishima, T. (1999). Realtime chord recognition of musical sound: a system using common lisp music. In *International Conference on Mathematics and Computing*.

Givens, C. R. and R. M. Shortt (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal 31*, 231–240.

Harte, C., M. B. Sandler, S. A. Abdallah, and E. Gómez (2005). Symbolic representation of musical chords: A proposed syntax for text annotations. In *International Society for Music Information Retrieval Conference*.

Mauch, M., S. Dixon, C. Harte, M. A. Casey, and B. Fields (2007). Discovering chord idioms through beatles and real book songs. In *International Society for Music Information Retrieval Conference*.

Mauch, M., K. C. Noland, and S. Dixon (2009). Using musical structure to enhance automatic chord transcription. In *International Society for Music Information Retrieval Conference*.

Oudre, L., Y. Grenier, and C. Févotte (2011). Chord recognition by fitting rescaled chroma vectors to chord templates. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 2222–2233.

Rocher, T., M. Robine, and P. Hanna (2010). A survey of chord distances with comparison for chord analysis. In *International Conference on Mathematics and Computing*.

Tymoczko, D. (2006). The geometry of musical chords. *Science 313*, 72 – 74.

Villani, C. (2009). *Optimal transport, old and new*, Volume 338. Springer Berlin, Heidelberg.

# Régularisation entropique décroissante pour le transport optimal semi-discret

Ferdinand Genans [1] & Antoine Godichon-Baggioni [2] & Olivier Wintenberger [3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, France, fgenans@lpsm.paris*
[2] *Laboratoire de Probabilités, Statistique et Modélisation, France, antoine.godichon_baggioni@sorbonne-universite.fr*
[3] *Laboratoire de Probabilités, Statistique et Modélisation, France, olivier.wintenberger@sorbonne-universite.fr*

**Résumé.** Le transport optimal est une méthode de comparaison des distributions de probabilité, utilisée dans diverses disciplines, y compris l'économie, l'apprentissage automatique et la biologie. Néanmoins, résoudre les problèmes de transport optimal est coûteux en calcul, ce qui a incité à introduire le transport optimal entropique. Cette dernière approche incorpore un terme de régularisation entropique pour faciliter des calculs plus abordables et efficaces. La sélection du paramètre de régularisation $\varepsilon$ devient alors une préoccupation pratique. Une régularisation plus faible est préférée pour la précision mais est souvent associée à une convergence plus lente, établissant un compromis entre le taux de convergence et la précision. Dans le cadre du transport optimal semi-discret, nous introduisons un algorithme de Descente de Gradient Stochastique, incorporant un schéma de régularisation décroissante.

**Mots-clés.** Transport Optimal, Optimisation Stochastique, Descente de Gradient, Régularisation Entropique.

**Abstract.** Optimal transport is a method for comparing probability distributions, used in various disciplines, including economics, machine learning, and biology. Nonetheless, solving optimal transport problems is computationally expensive, which has led to the introduction of entropic optimal transport. This latter approach incorporates an entropic regularization term to facilitate more affordable and efficient calculations. The selection of the regularization parameter $\varepsilon$ then becomes a practical concern. Weaker regularization is preferred for accuracy but is often associated with slower convergence, establishing a trade-off between convergence rate and accuracy. In the context of semi-discrete optimal transport, we introduce a Stochastic Gradient Descent algorithm, incorporating a decreasing regularization scheme.

**Keywords.** Optimal transport, Stochastic Optimization, Gradient Descent, Entropic Regularization.

# 1 Introduction

## 1.1 Transport Optimal

La théorie du transport optimal (TO) remonte aux travaux de Monge (1781) et a été généralisée plus tard par Kantorovich (1942). Ce cadre offre une méthode puissante pour comparer et transporter des mesures de probabilité et a démontré son efficacité dans divers domaines comme l'apprentissage automatique (Courty et al., 2014; Genevay et al., 2018; Bigot et al., 2017), la biologie (Schiebinger et al., 2019), et l'économie (Galichon, 2018).

Étant donné des mesures de probabilité source et cible $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ et une fonction $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ représentant le coût pour déplacer la masse, la formulation de Kantorovich du transport optimal est

$$\mathrm{TO}_c(\mu, \nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \mathrm{d}\pi(x, y). \tag{1}$$

Habituellement, nous utilisons $c(x, y) = \|x - y\|^p$ pour $p \geq 1$ et dans ce cas, $\mathrm{TO}_c^{\frac{1}{p}}$ représente une distance entre mesures de probabilité, appelée la distance de Wasserstein p (Villani, 2009; Santambrogio, 2015). Dans ce travail, nous nous concentrerons sur le coût $c(x, y) = \frac{1}{2}\|x - y\|^2$.

**Distance de Wasserstein 2 et théorème de Brenier.**

Soit $\mu \in \mathcal{P}(\mathbb{R}^d)$ une mesure source ayant une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ et une mesure cible $\nu \in \mathcal{P}(\mathbb{R}^d)$. De plus, supposons que $\mu$ et $\nu$ ont des moments d'ordre deux. Dans ce contexte, le théorème de Brenier (Brenier, 1991) stipule qu'il existe une carte unique $T_{\mu,\nu}$, que nous appelons la carte de Brenier, telle que

$$T^*_{\mu,\nu} := \operatorname*{argmin}_{T \in \mathcal{T}(\mu,\nu)} \int \frac{1}{2}\|x - T(x)\|^2 \mathrm{d}\mu(x), \tag{2}$$

où $\mathcal{T}(\mu, \nu) := \left\{T : \mathbb{R}^d \times \mathbb{R}^d, \mu(T^{-1}(B)) = \nu(B) \text{ pour tous } B \in \mathcal{B}(\mathbb{R}^d)\right\}$ est l'ensemble des cartes transportant $\mu$ sur $\nu$.

Cette carte résout également la formulation de Kantorovich du transport optimal dans le sens où la carte $(\mathrm{Id}, T_{\mu,\nu})_{\#}\mu$ résout (1) pour le coût $c(x, y) = \frac{1}{2}\|x - y\|^2$. De plus, $T^*_{\mu,\nu}$ est le gradient d'une fonction convexe et nous avons que

$$T^*_{\mu,\nu}(x) = x - \nabla f^*(x), \tag{3}$$

où $f^*$ est un potentiel de Kantorovich, c'est-à-dire, $(f^*, f^{*c})$ maximise le problème dual de Kantorovich

$$f^* \in \arg\max_{f \in C(\mathbb{R}^d)} \int_{\mathbb{R}^d} f \mathrm{d}\mu + \int_{\mathbb{R}^d} f^c \mathrm{d}\nu, \tag{4}$$

où nous

notons $f^c$ la $c$-transformation de $f$ donnée pour tout $y \in \mathbb{R}^d$ par

$$f^c(y) := \inf_{x \in \mathbb{R}^d} \left[ \frac{1}{2}\|x-y\|^2 - f(x) \right].$$

**Couplage de transport optimal entropique et formulation duale.**

L'ajout d'un terme entropique au transport optimal a été présenté dans les travaux de Cuturi (2013) pour le cas discret et est devenu populaire grâce à son calcul efficace via l'algorithme de Sinkhorn. Avec un $\varepsilon > 0$ fixé, le Transport Optimal Entropique (TOE) reformule Kantorovich comme

$$\mathrm{TO}_c^{\varepsilon}(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi \mid \mu \otimes \nu), \tag{5}$$

où

$$\mathrm{KL}(\pi \mid \mu \otimes \nu) \overset{\text{def.}}{=} \int_{\mathbb{R}^d \times \mathbb{R}^d} \ln\left( \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)} \right) \mathrm{d}\pi(x,y)$$

est l'entropie relative entre $\pi$ et $\mu \otimes \nu$. Comme mentionné dans Nutz and Wiesel (2022), le TOE a une solution unique $\gamma_{\mu,\nu}^{\varepsilon,*} \in \Pi(\mu,\nu)$ et a la densité suivante :

$$\frac{\mathrm{d}\gamma_{\mu,\nu}^{\varepsilon,*}}{\mathrm{d}(\mu \otimes \nu)}(x,y) = \exp\left( \frac{f_\varepsilon^*(x) + \mathbf{g}_\varepsilon^*(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon} \right), \tag{6}$$

pour deux fonctions mesurables $f_\varepsilon^* : \mathbb{R}^d \to \mathbb{R}$ et $\mathbf{g}_\varepsilon^* : \mathbb{R}^d \to \mathbb{R}$, parfois appelées les potentiels de Schrödinger. Ces potentiels sont uniques à une constante près ajoutée à une fonction et soustraite à l'autre (Nutz and Wiesel, 2022) et solution du problème dual du TOE

$$\sup_{f \in L^1(\mu), g \in L^1(\nu)} \int f \mathrm{d}\mu + \int g \mathrm{d}\nu - \varepsilon \int \exp\left( \frac{f(x) + g(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon} \right) \mathrm{d}\mu(x)\mathrm{d}\nu(y) + \varepsilon. \tag{7}$$

En notant $(f^*, f^{*c})$, résolvant (4), Carlier et al. (2017) fournit la convergence de $(f_\varepsilon^*, g_\varepsilon^*)$ vers $(f^*, f^{*c})$ lorsque $\varepsilon \to 0$.

# 2 Cadre: transport optimal semi-discret

Dans ce travail, nous nous intéressons au TO semi-discret, cas où la mesure source $\mu$ est continue, tandis que la mesure cible $\nu$ est discrète. Nous considérons les hypothèses suivantes, présentes dans Delalande (2022) et Divol et al. (2022).

**Hypothèses.** ***(A)*** *La mesure source $\mu$ est une mesure continue avec* $\mathrm{Supp}(\mu) \subset B(0,R)$ *et une densité $p$ satisfaisant* $0 < p_{\min} \leq p \leq p_{\max} < \infty$ *pour certaines constantes* $p_{\min}, p_{\max}$ *et* $R$.

***(B)*** *La mesure cible $\nu$ est une mesure discrète de la forme*

$$\nu = \sum_{j=1}^{M} b_j \delta_{\mathbf{y_j}}$$

*avec* $\mathbf{b} = (b_1, \ldots, b_M)$ *ses poids de probabilité et* $(\mathbf{y_1}, \ldots, \mathbf{y_M}) \in B(0, R)^M$ *son support.*

**Formulation semi-duale.**

Supposant que nous pouvons échantillonner à partir de la mesure source $\mu$, Genevay et al. (2016) a suggéré d'utiliser une formulation semi-duale pour résoudre le problème du TOE semi-discret. Cela aboutit à un problème d'optimisation convexe lisse consistant à trouver un minimiseur de la fonction $H_\varepsilon$ définie pour tout $\mathbf{g} = (g_1, \ldots, g_M)$ par

$$H_\varepsilon(\mathbf{g}) \stackrel{\text{def.}}{=} -\int_{\mathbb{R}^d} \mathbf{g}^{c,\varepsilon}(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) - \sum_{j=1}^{M} g_j b_j, \tag{8}$$

où pour $\mathbf{x} \in \mathbb{R}^d$, $g^{c,\varepsilon}(\mathbf{x})$, souvent désigné comme la e $(c, \varepsilon)$-transformée (Peyré et al., 2019) est défini par

$$\mathbf{g}^{c,\varepsilon}(x) := -\varepsilon \ln \left( \sum_{j=1}^{M} \exp\left( \frac{g_j - \frac{1}{2}\|x - y_j\|^2}{\varepsilon} \right) b_j \right).$$

Notons que la formulation (8) permet de ne pas avoir un biais de discrétisation, comme on aurait eu en échantillonnant directement $\mu$ pour appliquer l'algorithme de Sinkhorn Cuturi (2013). Nous pouvons calculer le coût du TOE avec (3) puisque nous avons $\min_{\mathbf{g} \in \mathbb{R}^M} H_\varepsilon(\mathbf{g}) = \mathrm{OT}_c^\varepsilon(\mu, \nu)$ (Genevay et al., 2016). L'un des intérêts de cette formule est que la fonctionnelle $H_\varepsilon$ est différentiable et son gradient est défini pour tout $\mathbf{g} \in \mathbb{R}^M, \mathbf{x} \in \mathbb{R}^d$ et toute coordonnée $j$ par

$$\nabla_{\mathbf{g}} H_\varepsilon(\mathbf{g}, \mathbf{x})_j = -b_j + \int_{\mathbb{R}^d} \chi_j^\varepsilon(\mathbf{x}, \mathbf{g}) \mathrm{d}\mu(\mathbf{x}),$$

où pour $\mathbf{x} \in \mathbb{R}^d, \mathbf{g} \in \mathbb{R}^M$, nous avons

$$\chi_j^\varepsilon(\mathbf{x}, \mathbf{g}) = \frac{\exp\left( \frac{-c(\mathbf{x}, \mathbf{y}_j) + g_j}{\varepsilon} \right) b_j}{\sum_{k=1}^{M} \exp\left( \frac{-c(\mathbf{x}, \mathbf{y}_k) + g_k}{\varepsilon} \right) b_k}.$$

**Résoudre le semi-discret avec l'optimisation stochastique et en ligne.**

Étant donné une mesure source arbitraire dont nous pouvons échantillonner, la formulation semi-duale est particulièrement pertinente pour l'estimation du potentiel de Brenier. En effet, cette formulation nous permet de ne pas discrétiser la mesure source et ainsi de surmonter un biais de discrétisation qui aurait été obligatoire pour utiliser la Programmation Linéaire

ou les algorithmes de Sinkhorn. On peut considérer le problème de minimisation donné par (8) comme un problème d'Optimisation Stochastique. En effet, la fonctionnelle $H_\varepsilon$ peut être réécrite comme

$$H_\varepsilon(\mathbf{g}) = \mathbb{E}\left[h_\varepsilon(X, \mathbf{g})\right] = \int_{\mathbb{R}^d} h_\varepsilon(\mathbf{x}, \mathbf{g}) \mathrm{d}\mu(\mathbf{x}),$$

où $X$ représente une variable aléatoire tirée de la distribution source $\mu$, et pour tout $(\mathbf{x}, \mathbf{g}) \in \mathbb{R}^d \times \mathbb{R}^M$

$$h_\varepsilon(\mathbf{x}, \mathbf{g}) = -\mathbf{g}^{c,\varepsilon}(\mathbf{x}) - \sum_{j=1}^{M} g_j b_j.$$

Pour tout $\mathbf{g} \in \mathbb{R}^M$, étant donné $\mathbf{x} \sim \mu$, un estimateur non biaisé du gradient est alors, pour toute coordonnée $1 \leq j \leq M$

$$\nabla_{\mathbf{g}} h_\varepsilon(\mathbf{g}, \mathbf{x})_j = -b_j + \chi_j^\varepsilon(\mathbf{x}, \mathbf{g}).$$

Pour une régularisation fixe $\varepsilon > 0$, les méthodes stochastiques de premier ordre sont principalement employées pour ce cadre. Spécifiquement, dans Genevay et al. (2016), le Descente de Gradient Stochastique Moyennée (ASGD) est utilisée. Ayant une initialisation $\mathbf{g}_1 \in \mathbb{R}^M$, et fixant $\gamma_t = \gamma_0/t^b$ avec $b \in (0, 1)$, ASGD consiste à chaque itération à considérer un bloc de $n \geq 1$ données i.i.d $\mathbf{x}_t = (\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,n}) \sim \mu^{\otimes n}$ et ensuite faire la mise à jour

$$\mathbf{g}_{t+1} = \mathbf{g}_t - \gamma_t \nabla_{\mathbf{g}} h_\varepsilon(\mathbf{x}_t, \mathbf{g}_t), \tag{9}$$

$$\bar{\mathbf{g}}_{t+1} = \frac{1}{t+2}\mathbf{g}_{t+1} + \frac{t+1}{t+2}\bar{\mathbf{g}}_t \tag{10}$$

où $\nabla_{\mathbf{g}} h_\varepsilon(\mathbf{x}_t, \mathbf{g}_t) := \frac{1}{n}\sum_{i=1}^n \nabla_{\mathbf{g}} h_\varepsilon(\mathbf{x}_t, \mathbf{g}_{t,i})$.. Nous parlons de moyennisation puisque $\bar{\mathbf{g}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{g}_t$. A noter que les algorithmes ASGD sont largement étudiés (Polyak and Juditsky, 1992; Pelletier, 2000; Bach and Moulines, 2013) et on peut se référer Bercu and Bigot (2021) pour le cas spécifique du TOE.

# 3 Régularisation entropique décroissante

Comme nous l'avons vu, la régularisation entropique permet d'obtenir une fonction objectif différentiable et lisse, facilitant ainsi l'utilisation des méthodes d'optimisation basées sur le gradient. Cependant, le choix du coefficient de régularisation $\varepsilon$ reste critique. Plus $\varepsilon$ est faible, plus l'algorithme se rapproche du problème non régularisé. Sa convergence, cependant, ralentit en fonction de $\varepsilon$, établissant un équilibre entre la vitesse de convergence et la précision de l'approximation. Pour approximer le TO non régularisé, notamment le Transport de Brenier, l'utilisation d'une régularisation décroissante au fil du temps, notée $\varepsilon_t$, semble naturel dans un cadre stochastique ou en ligne. En se basant sur le travail récent de Delalande (2022) sur la différence d'optimiseur entre deux régularisations différentes lorsque nos mesures vérifient les hypothèses (A) et (B), nous introduisons dans cette section un algorithme de Descente de Gradient Stochastique avec régularisation décroissante.

**Étape de projection.**

Dans la descente de gradient stochastique et en ligne, l'incorporation d'une étape de projection avec une connaissance préalable de l'espace des solutions peut conduire à une convergence plus rapide. Ceci est particulièrement vrai lorsque l'étape de projection est presque sans coût computationnel. Nous introduisons ici deux ensembles de projection différents qui peuvent être utilisés pour les problèmes OT et EOT, en tirant parti de la régularité des solutions dans (8). À cette fin, remarquons que pour toute fonction de coût bornée dans $\mathrm{Supp}(\mu) \times \mathrm{Supp}(\nu)$ et pour toute régularisation $\varepsilon > 0$, il existe une solution optimale $\mathbf{g}_\varepsilon^* = (g_1^*, \ldots, g_M^*)$ de la formulation semi-duale telle que pour tous $j = 1, \ldots, M$ nous avons $0 \leq g_j^* \leq \|c\|_\infty$, où nous notons $\|c\|_\infty$ le supremum de la fonction de coût sur $\mathrm{Supp}(\mu) \times \mathrm{Supp}(\nu)$ (voir Nutz and Wiesel (2022) par exemple). Nous pouvons donc également ajouter une étape de projection sur l'ensemble convexe compact suivant

$$C_\infty := [0, \|c\|_\infty]^M.$$

Cette projection est presque sans coût puisqu'elle consiste uniquement à clipper chaque coordonnée de notre vecteur. Cependant, pour certains coûts, nous pouvons trouver un meilleur ensemble de projection, en utilisant la régularité de la fonction de coût. Cet ensemble de projection est donné par le lemme suivant.

**Lemme 1.** *Si pour tous $\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \mathbb{R}^d$, il existe des constantes $K, \beta$ telles que*

$$|c(\mathbf{x}, \mathbf{y}) - c(\mathbf{x}, \mathbf{y}')| \leq K\|\mathbf{y} - \mathbf{y}'\|^\beta,$$

*alors, pour tous $k = 1, \ldots, M$, il existe une solution unique $\mathbf{g}^*$ à (3) dans l'ensemble convexe compact suivant*

$$C_k := \{\mathbf{g} \in \mathbb{R}^M;\, g_k = 0 \text{ et } |g_j| \leq K\|\mathbf{y_k} - \mathbf{y_j}\|^\beta,\ j = 1, \ldots, M\}\ .$$

**Algorithme ASGD projeté avec régularisation décroissante.**

Tirant parti de notre étape de projection, nous proposons de changer l'étape usuelle de descente de gradient avec régularisation fixe, par une régularisation décroissante $\varepsilon_t$ avec $\varepsilon_t \to 0$, ce qui conduit à l'algorithme de gradient projeté suivant

$$\mathbf{g}_t = \Pi_C \left( \mathbf{g}_{t-1} - \frac{\gamma_1}{t^b} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{x}_t, \mathbf{g}_{t-1}) \right).$$

Ce mécanisme de régularisation offre l'avantage d'une forte régularisation en début d'algorithme, qui s'affine progressivement pour améliorer la précision de notre estimateur au fil du temps, visant ainsi à obtenir une estimation précise et non biaisée du potentiel de Brenier. La version complète de l'algorithme est décrite ci-dessous.

**Algorithm 1** Decreasing Regularization Projected ASGD (DRPASGD)

**Parameters:** $(\gamma_1, \gamma, a, b, n)$
Initialize $\mathbf{g}_0 \in C$ and $\bar{\mathbf{g}}_0 = \mathbf{g}_0$
**for** $t = 0$ to $T-1$ **do**
$\varepsilon_t = 1/t^a$
$\gamma_t = \gamma_1/t^b$
$x_t \sim \mu^{\otimes n}$
$\mathbf{g}_{t+1} = \mathrm{Proj}_C\left(\mathbf{g}_t - \gamma_{t+1}\frac{1}{n}\sum_{k=1}^{n} \nabla_{\mathbf{g}} h_{\varepsilon_t}(x_{t,i}, \mathbf{g}_t)\right)$
$\bar{\mathbf{g}}_{t+1} = \dfrac{1}{t+2}\mathbf{g}_{t+1} + \dfrac{t+1}{t+2}\bar{\mathbf{g}}_t$
**end for**
**return** $\bar{\mathbf{g}}_T$

## 3.1 Illustrations numériques

**Convergence de l'estimateur.**

Nous considérons une expérience synthétique pour l'illustration de la convergence de DRPASGD. Soit $\mu = \mathrm{Unif}\left([0,1]^{10}\right)$ , et fixons $M = 100$. Nous générons aléatoirement $y_1, \ldots, Y_{100} \in [0,1]^{10}$, ainsi qu'un potentiel $\mathbf{g}^* \in [0,1]^{100}$ aléatoire, et considérons le transport optimal

$$T_0(x) = \mathrm{argmin}_{j \in [\![1,M]\!]} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{y}_j\|2 - g_j^* \right\}.$$

Nous définissons $\nu = (T_0)_\sharp P$, comme cela $\mathbf{g}^*$ est bien un potentiel optimal, et approximons les poids $\mathbf{b}$ par Monte-Carlo avec $10^6$ échantillons de $\mu$.

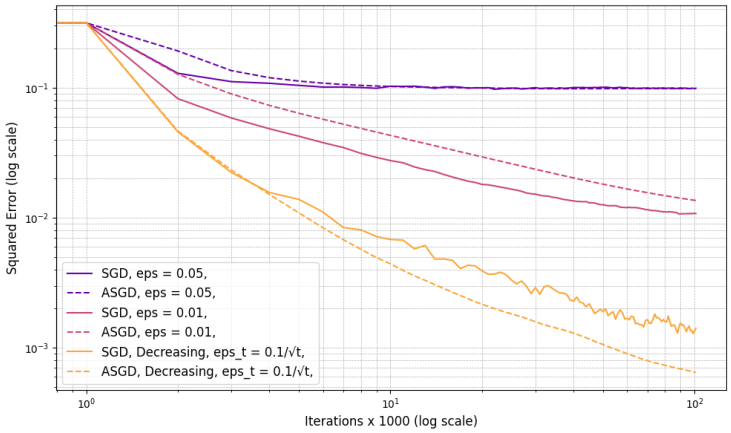


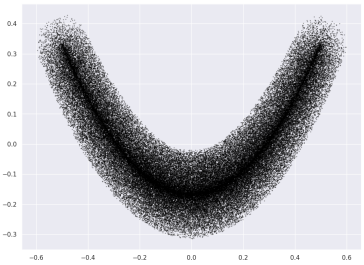

Figure 1: Évolution moyenne de l'erreur $\|\mathbf{g}_t - \mathbf{g}^*\|_2^2$ pour chaque estimateur, $t \in [0, 10^5]$, sur 20 expériences par configuration.

Nous illustrons dans la Figure 1 l'utilité de DRPASGD, en fixant $\gamma_t = \frac{5}{t^{3/4}}$ pour ASGD et DRPASGD, pour les deux algorithmes, nous utilisons la projection sur $[0,1]^{100}$ et choisissons

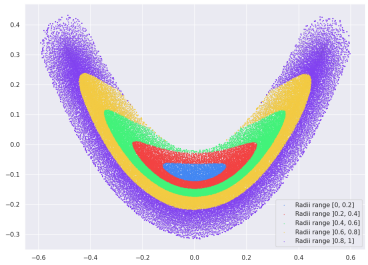
une régularisation décroissante $\varepsilon_t = 0.1/\sqrt{t}$. On constate que DRPASGD bénéficie d'une accélération tout en ne s'arrêtant pas à un optimum biaisé contrairement aux régularisations fixes.

**Quantiles de Monge-Kantorovich.**

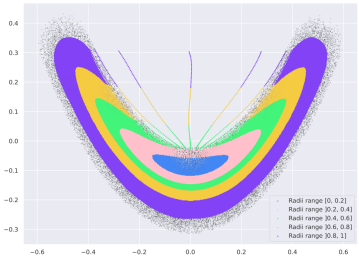
Nous illustrons ici notre méthode pour déterminer les régions quantiles de Monge-Kantorovich Chernozhukov et al. (2017), qui définissent une généralisation en dimension $d > 1$ de la notion habituelle de quantiles, en s'appuyant sur le théorème de Brenier Brenier (1991). Dans ce cas, la mesure source $\mu$ est la mesure uniforme sur la boule unité euclidienne. Nous fixons ici $d = 2$ et prenons comme mesure discrète cible,discrétisation d'une mesure en forme de "banane" avec $M = 10^5$ points, exemple que l'on peut retrouver dans Chernozhukov et al. (2017) et Bercu et al. (2023). Nous illustrons le résultat obtenu par DRPASGD et le comparons avec une régularisation fixe de $\varepsilon = 0.002$. En commençant au centre, chaque région de couleur correspond à la région quantile $[0.2k, 0.2(k+1)]$ où $k \in [0.4]$, c'est à dire, là où les points de $\mathcal{B}(0, 0.2(k+1)) \setminus \mathcal{B}(0, 0.2k)$ sont envoyés.



(a) Mesure discrète cible, $M = 10^5$.



(b) MK-régions, DRPASGD, $\varepsilon_t = 0.1/\sqrt{t}$.



(c) MK-régions, transport entropique, $\varepsilon = 0.002$.

Figure 2: MK-Quantiles pour notre distribution "banane", après $T = 10^6$ itérations.

Comme on le voit sur la figure 2, notre algorithme permet d'avoir des régions quantiles recouvrant toute la distribution, tandis qu'avec une régularisation fixe, le dernier contour ne contient pas toute la distribution.

# Bibliographie


F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

B. Bercu and J. Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. 2021.

B. Bercu, J. Bigot, and G. Thurin. Stochastic optimal transport in banach spaces for regularized estimation of multivariate quantiles. *arXiv preprint arXiv:2302.00982*, 2023.

J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic pca in the wasserstein space by convex pca. 2017.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–kantorovich depth, quantiles, ranks and signs. 2017.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances In Neural Information Processing Systems*, 26, 2013.

A. Delalande. Nearly tight convergence bounds for semi-discrete entropic optimal transport. In *International Conference On Artificial Intelligence And Statistics*, pages 1619–1642, 2022.

V. Divol, J. Niles-Weed, and A.-A. Pooladian. Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*, 2022.

A. Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances In Neural Information Processing Systems*, volume 29, 2016.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

M. Nutz and J. Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1-2):401–424, 2022.

M. Pelletier. Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM Journal on Control and Optimization*, 39(1):49–72, 2000.

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

F. Santambrogio. Optimal transport for applied mathematicians. 2015.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

# Statistique appliquée à la médecine 1

# Statistical analysis of matched survival data in national health databases.

Vanessa CHEZEU[1] & Jean-Francois DUPUY[2]& Valerie GARES[3]
& Samuel BOWONG[4]& Andre NANA[5]

[1] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France/ Univerty of Douala, Cameroun; Fidelette.Chezeu-Toumeni@insa-rennes.fr*
[2] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France; Jean-Francois.Dupuy@insa-rennes.fr*
[3] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, France; Valerie.Gares@insa-rennes.fr*
[4] *University of Douala, Cameroun; sbowong@gmail.com*
[5] *University of Douala, Cameroun; nanayakam@yahoo.com*

**Résumé.**

Nous nous intéressons à l'estimation des paramètres du modèle à risques proportionnels de Cox, à partir de bases de données de santé appariées. Nous considérons la situation dans laquelle les variables explicatives et les durées de vie des individus ne sont pas rapportées dans la même base de données. Un processus préalable de couplage probabiliste d'enregistrements (chaînage) est donc nécessaire pour obtenir une base de données complète. Dans ce travail, nous proposons une équation d'estimation des paramètres du modèle de Cox, adaptée à ce cadre. Cette équation est obtenue en adaptant la fonction de score partiel "usuelle" du modèle de Cox, afin de prendre en compte le processus préalable de couplage des données. Nous décrivons une première étude de simulation, dans laquelle les probabilités d'appariement de chaque paire couplée sont déjà disponibles. Au travers de cette étude, nous évaluons les propriétés des estimateurs proposés. Dans une seconde approche, nous simulons deux bases de données, puis estimons les probabilités d'appariement des individus de ces bases (à l'aide d'un modèle "recordlinkage"), avant d'appliquer la méthodologie d'estimation proposée. Nous décrivons également les résultats de cette étude.

**Mots clé.** Couplage d'enregistrements ; Données censurées ; Durées de vie ; Santé publique ; Simulations numériques.

**Abstract.**

In this work, we investigate estimation in the Cox proportional hazards model from matched health databases. We consider the situation where the explanatory variables and individual lifetimes are not reported in the same database. A prior process of probabilistic record linkage is therefore necessary to obtain a complete database. We propose an estimating equation for the Cox model, adapted to this framework. This equation is obtained by adapting the "usual" partial score function, in order to take account of the prior linkage data process. We assess the properties of the resulting estimate via simulations. In the first simulation study, we assume that the matching probabilities of each linked pair are already available. In a second study, we simulate two databases, then we estimate the matching

probabilities of their respective individuals (using a record linkage model), and we apply the proposed estimation methodology. Results are described.

**Keywords.** Record linkage ; Censored data ; Duration data ; Public health ; Numerical simulations.

# 1 Introduction

Survival analysis refers to the set of statistical methods used to analyse data where the outcome of interest is the time of occurrence of some event (such as death, relapse...). Survival data occur in a wide range of fields: economy (unemployment duration), finance (time until repayment of a loan), insurance (duration between the beginning of long term care and death), engineering and reliability (duration until failure of an engine). In this work, we consider the statistical analysis of survival data arising from matched health databases.

The National Health Data System ('SNDS') is a large health database, which is often used to enrich existing medical cohorts and registries. This enrichment allows to recover as much information as possible on the evolution of patients health status. This, in turn, allows to make more robust analysis, but taking advantage of complementary data.

The enrichment of health databases can be done through a linkage data process between two databases. This linkage is simple if one has access to some unique patients identifiers (such as a unique code assigned definitively to a patient from his first contact within the establishment). However the use of this identifier may not be permitted for ethical reasons, or an identifier may simply not be available. In this case, we may only use partial identifiers, which are common to these databases (such as gender, postal code, dates of treatment...) to identify matched pairs from both databases.

Probabilistic record linkage method was first developed by Fellegi and Sunter (1969). Record linkage is a process of combining information about an individual or event in two or more databases. That is, the data from one source is joined with the data from another source that describes the same entity. For each pair of records, this method provides a score (or matching probability) which makes it possible to take into account, in the statistical analysis, the errors related to the matching process. To date, there have been several applications and improvements of the Fellegi and Sunter method, see for example Thanh et al. (2022) and Danhyang et al. (2022). Several authors have considered survival analysis from linked databases. For example, Thanh et al. (2023) propose an estimating equation in Cox proportional hazards (PH) model when no information on matching variables is available to the analyst, and the linkage errors are estimated from a validation file.

In the present work, we consider the situation where two databases are available: one of them contains the patients survival times (and censoring information), the second one contains information on the patients covariates. Matching probabilities between pairs of patients in both databases are available. We propose some estimation methods in the PH model, adapted

to this setting, which has yet received little attention (see Lahiri and Larsen (2005), Hof et al. (2017), Ying and Partha (2019)).

Our work is structured as follows: in section 2, we describe our problem. In section 3, we define the record linkage model. In section 4, we formulate the model and we propose an estimating equation for the parameters of the Cox model, based on the linked data. In section 5, we assess, via numerical simulations, the properties of the proposed estimate.

# 2 Problem description

Let us consider two databases $A$ and $B$ with respective sizes $n_A$ and $n_B$ such that $n_A \leq n_B$, and all individuals of database $A$ are included in database $B$. The database $B$ contains information on $n_B$ independent individuals, such that for each individual $j = 1, \ldots, n_B$, we observe a pair $\{\boldsymbol{X}_j, \boldsymbol{Y}_j\}$, where $\boldsymbol{X}_j = (X_j^1, X_j^2, \ldots, X_j^P)^\top$ is a vector containing measures of the $P$ covariates (e.g. blood group, monthly income), and $\boldsymbol{Y}_j = (Y_j^1, Y_j^2, \ldots, Y_j^K)^\top$ is a vector containing the measures of $K$ variables, considered as partial identifiers of individual $j$ (e.g. name, age, postal code). These variables are usually called matching variables.
The database $A$ contains survival data of $n_A$ independent individuals, such that for each individual $i = 1, \ldots, n_A$, we observe the triplet $\{T_i, \delta_i, \boldsymbol{Y}_i\}$, where $\boldsymbol{Y}_i$ is the same $K$-vector of partial identifier variables as in database $B$, $T_i = \min(\tilde{T}_i, C_i)$ is the observed survival time, $\tilde{T}_i$ is the event time of interest (e.g., time between inclusion in a trial and death), $C_i$ is a random right-censoring time, and $\delta_i = 1_{\tilde{T}_i \leq C_i}$ is the censoring indicator. The event of interest is observed for an individual $i$ if it occurs before the censoring time $C_i$. If for an individual $i$, we did not observe the event before $C_i$, this individual is considered to be censored (we have no information about its state after this period).

We assume that a unique individual $j$ in $B$ corresponds to each individual $i$ in $A$, and that the record linkage errors are non-informative of the regression model (that is, matching errors can depend only on errors in the matching process, but not on covariates, nor on survival time data, see Thanh Vo and al (2022)).

Our objective is to assess the relationship between the patients lifetimes (which are recorded in database $A$) and the covariates (which are recorded in the database $B$ but not in $A$).

# 3 Probabilistic record linkage model

In order to constitute a complete health database which contains information from both databases $A$ and databases $B$, it is necessary to identify which measurements from databases $A$ and databases $B$ belong to the same individual. Due to the absence of unique identifiers in most situations, we must use the partial identifier variables $\boldsymbol{Y} = (Y^1, Y^2, \ldots, Y^K)^\top$ common to both databases.

Let, $\Omega = A \times B = \{(i, j) : i \in A \text{ and } j \in B; i = 1, \ldots, n_A; j = 1, \ldots, n_B\}$ the space which contains all the comparison pairs of units from $A$ and $B$. For each pair of individuals $(i, j) \in$

$\Omega$, the value of the matching variables are compared to each other. In this context, we use the binary comparison method proposed by Fellegy and Sunter (1969). Let the comparison function for each variable $Y^k$ defined as,

$$\forall k = \{1, \ldots, K\}, \Gamma_{ij}^k = \begin{cases} 1 \text{ if } Y_i^k = Y_j^k, \\ \\ 0 \text{ if } Y_i^k \neq Y_j^k. \end{cases}$$

For each pair $(i, j) \in \Omega$, one obtains a comparison vector,

$$\mathbf{\Gamma}_{ij} = \left(\Gamma_{ij}^1, \ldots, \Gamma_{ij}^k, \cdots, \Gamma_{ij}^K\right) = (\Gamma_{ij}^k)_{1 \leq k \leq K}.$$

According to Fellegi and Sunter, every possible pair of individuals belongs either to the "matched" set $M$ or to the "unmatched" set $U$ such that,

$$M = \{(i, j); i = j, i \in A, j \in B\}, \quad \text{and} \quad U = \{(i, j); i \neq j, i \in A, j \in B\}.$$

If there is no error in the partial identifier variables, only the record pairs which have a comparison vector $\mathbf{\Gamma}_{ij} = (1, \ldots, 1)$ will be observed in the set $M$. Also, if the partial identifier variables were unique for each individual, only pairs which have a comparison vector $\mathbf{\Gamma}_{ij} = (0, \ldots, 0)$ will be observed in the set $U$. Because of errors generally present in databases (often due to bad recording of information, error in coding, transcription), we could observe pairs in the set $M$ with different comparison vectors (e.g. $\mathbf{\Gamma}_{ij} = (0, 1, \ldots, 1)$).

Considering each comparison vector $\mathbf{\Gamma}_{ij}$, the probabilistic record linkage associates to each pair a probability of belonging to one of the two subsets of $\Omega$.

Let $\boldsymbol{\gamma}_{ij} = \left(\gamma_{ij}^1, \ldots, \gamma_{ij}^K\right) \in \{0; 1\}^K$ be the realization set of the random variable $\mathbf{\Gamma}_{ij}$. The distribution of the comparison vector $\mathbf{\Gamma}_{ij}$ for each pair $(i, j)$ is given by the following model:

$$\mathbb{P}\left(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij}\right) = \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in M)\mathbb{P}((i, j) \in M) + \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in U)\mathbb{P}((i, j) \in U). \tag{1}$$

In practice, the number of distinct values of $\mathbf{\Gamma}_{ij}$ can be large, and estimation of the probabilities becomes complicated. Fellgi and Sunter (1969) and some authors (Andersen, and Gill (1982)) have therefore consider the hypothesis that the components of the vector $\mathbf{\Gamma}_{ij}$ can be reorganized, and are mutually statistically independent. According to the independence between the components of the vector $\boldsymbol{\gamma}_{ij}$, one has $\forall i = 1, \cdots n_A; j = 1, \cdots, n_B$:

$$\mathbb{P}\left(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in M\right) = \prod_{k=1}^{K} \mathbb{P}(\Gamma_{ij}^k = \gamma_{ij}^k \mid (i, j) \in M),$$

and

$$\mathbb{P}\left(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in U\right) = \prod_{k=1}^{K} \mathbb{P}(\Gamma_{ij}^k = \gamma_{ij}^k \mid (i, j) \in U),$$

Let us define by: $m^k = \mathbb{P}(\Gamma_{ij}^k = 1 \mid (i,j) \in M)$ , $u^k = \mathbb{P}(\Gamma_{ij}^k = 1 \mid (i,j) \in U)$ and $\pi_M = \mathbb{P}((i,j) \in M); \theta = (u^k, m^k, \pi_M; k = 1, \cdots, K)$ the set of all parameters to be estimated. We have a total of $(2K+1)$ parameters.

The final objective is to estimate $q_{ij} = \mathbb{P}((i,j) \in M \mid \mathbf{\Gamma}_{ij})$ ( the probability for having match knowing the comparison vector) and $\mathbb{P}((i,j) \in U \mid \mathbf{\Gamma}_{ij}) = 1 - q_{ij}$ .

Let be the matrix $\mathbf{\Gamma} = \{\boldsymbol{\gamma}_{ij}; i = 1, \cdots n_A; j = 1, \cdots n_B\}$. Considering the independence hypothesis for all comparison vectors, the likelihood function bored on $\mathbf{\Gamma}$ is defined by:

$$L(\theta \mid \mathbf{\Gamma})) = \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} \left[\pi_M \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in M) \mathbb{P}((i,j) \in M)\right]^{z_{ij}} \times \left[\pi_U \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in U) \mathbb{P}((i,j) \in U)\right]^{1-z_{ij}},$$

where, $\pi_U = 1 - \pi_M$, and $z_{ij}$ the indicator function such that:

$$z_{ij} = \begin{cases} 1 \text{ if } (i,j) \in M, \\ \\ 0 \text{ otherwise .} \end{cases}$$

The problem is then to maximize $L(\theta \mid \mathbf{\Gamma})$ under the constraint $\pi_U + \pi_M = 1$.

The EM algorithm (Meng and Rubin ( 1993), F.Santos (2015)) consists in estimating the parameters $\theta$ from a set of incomplete data. This algorithm makes it possible to determine the parameter $\theta$ which maximizes the likelihood function of the considered model.

Once all parameters are estimated using EM algorithm, the posterior probabilities are estimated for all pair $(i,j)$ by the Bayesian formula

$$\begin{aligned} q_{ij} &= \mathbb{P}((i,j) \in M \mid \mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij}) \\ &= \frac{\pi_M \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in M)}{\pi_M \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in M) + (1 - \pi_M) \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in U)} \end{aligned} \tag{2}$$

# 4 Cox regression with linked data

## 4.1 Cox proportional hazards model

Cox proportional hazards (PH) model (1972) is the most widely used survival regression model. It allows to estimate the effect of covariates $\boldsymbol{X}$ on patients lifetimes $T$. It is defined through the conditional hazard function:

$$\lambda(t \mid \boldsymbol{X}) = \lim_{dt \to 0} \frac{\mathbb{P}\left[t \leq T < t + dt \mid T \geq t, \boldsymbol{X}\right]}{dt}.$$

In Cox PH model, the conditional hazard function is specified as:

$$\lambda(t \mid \boldsymbol{X_i}) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i), \tag{3}$$

where $\boldsymbol{X}_i = (X_i^1, \cdots, X_i^p)^\top$ is a vector of covariates for individual $i$, $\lambda_0(t)$ is an unknown non-negative function of time (the so-called baseline hazard function) and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is a $p$-vector of unknown parameters to be estimated.

An estimator of $\beta$ is obtained by maximizing the partial likelihood, given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left( \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i)}{\sum_{j=1}^{n} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)} \right)^{\delta_i}, \tag{4}$$

where $Y_j(t) = 1_{T_j \geq t}$ is the indicator that individual $j$ is still at risk at time $t$. Differentiating $\log L(\boldsymbol{\beta})$ with respect to $\beta$ yields the following estimating equation:

$$H_{\text{Cox}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left( \boldsymbol{X}_i - \frac{\sum_{j=1}^{n} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j) \boldsymbol{X}_j}{\sum_{j=1}^{n} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)} \right) = 0.$$

The solution of this equation is called the maximum partial likelihood estimator of $\beta$. It is consistent and asymptotically normal, see Andersen and Gill (1982).

## 4.2 Estimation with linked data

We wish to estimate $\boldsymbol{\beta}$ in model (3), based on databases $A$ and $B$ described in section 2. Let $i$ be some individual in database $A$. Recall that we do not observe $\boldsymbol{X}_i$. We only know that the covariate vector for individual $i$ takes one value from the set $\{\boldsymbol{X}_1, \cdots \boldsymbol{X}_{n_B}\}$ in $B$.
Let us denote by $\boldsymbol{Z}_i$ the covariate vector that will be affected to individual $i$. From the record linkage process described above, we only know that

$$\mathbb{P}(\boldsymbol{Z_i} = \boldsymbol{X_j}) = q_{ij}, \quad j = 1, \ldots, n_B,$$

where the $\{q_{ij},\ i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B\}$ are defined by (2). We normalize the $q_{ij}$ by

$$\frac{q_{ij}}{\sum_{j=1}^{n_B} q_{ij}}$$

so that they sum to 1. We propose several estimation methods for $\boldsymbol{\beta}$.

**Method 1: a naive approach.** A first idea is to affect, to every individual $i$, the covariate vector $\boldsymbol{X}_j$ in $B$ which has the largest posterior probability $q_{ij}$, that is:

$$\boldsymbol{Z}_i = \boldsymbol{X}_j \quad \text{where} \quad j = \text{argmax}_{1 \leq j \leq n_B}(q_{ij})$$

By naively treating the linked covariates $\boldsymbol{Z}_i$ as the true covariates $\boldsymbol{X}_i$, the partial likelihood (4) becomes:

$$L_{\text{naive}}(\boldsymbol{\beta}) = \prod_{i=1}^{n_A} \left( \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_i)}{\sum_{j=1}^{n_B} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_j)} \right)^{\delta_i},$$

from which we can estimate $\boldsymbol{\beta}$.

**Method 2: a weighted partial likelihood.** In order to reduce the bias of the naive estimator in method 1, we propose to modify the partial likelihood, by taking into account the probabilistic aspect of the linked covariate $\boldsymbol{Z}_i$. The basic idea is as follows.
Consider the $i$-th individual in database $A$. We successively affect each of the $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_B}\}$ to $\boldsymbol{Z}_i$, thus creating $n_B$ fictional individuals. Each of these individuals enters the partial likelihood, but is suitably weighted by the matching probability of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. This yields the following weighted partial likelihood:

$$L_{\text{weighted}}(\boldsymbol{\beta}) = \prod_{i=1}^{n_A} \left( \prod_{j=1}^{n_B} \left[ \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)}{\sum_{p=1}^{n_A} Y_p(T_i) \left( \sum_{r=1}^{n_B} \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_r) q_{pr} (\tilde{q}_p^{-1}) \right)} \right]^{q_{ij}(\tilde{q}_i^{-1})} \right)^{\delta_i},$$

where $\tilde{q}_i = \min_{1 \leq j \leq n_B} q_{ij}$, and $\tilde{q}_p = \min_{1 \leq j \leq n_B} q_{pj}$.

Maximizing $L_{\text{weighted}}$ provides a second estimate of $\boldsymbol{\beta}$.

**Method 3: complete partial likelihood with unobserved variables**. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_B}\}$ be the set of possible values of $\boldsymbol{Z}_i$. If $\boldsymbol{Z}_i$ were observed, we could write the following likelihood, based on the observations $\{t_i, \delta_i, \boldsymbol{Z}_i\}, i = 1, \ldots, n_A$:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n_A} \mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i\right).$$

In fact, $\boldsymbol{Z}_i$ is not observed, therefore, we propose to maximize the conditional expectation of the log-likelihood for complete data, given the observations. This is the idea of the EM algorithm. We have:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}|T,\delta} \left( \log \left( L(\boldsymbol{\beta}) \mid (T, \delta) \right) \right) &= \sum_{i=1}^{n_A} \mathbb{E}_{\boldsymbol{Z}|T,\delta} \left( \log \left( \mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i\right) \right) \mid (t_i, \delta_i)_{1 \leq i \leq n_A} \right), \\
&= \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \log \left( \mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \right) \left( \mathbb{P} \left( \boldsymbol{Z}_i = \boldsymbol{x}_j \mid (t_i, \delta_i) \right) \right),
\end{aligned}$$

where

$$\log \left( \mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \right) = \log \left[ \mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \mathbb{P} \left( \boldsymbol{Z}_i = \boldsymbol{x}_j \right) \right],$$

$$\mathbb{P} \left( \boldsymbol{Z}_i = \boldsymbol{x}_j \mid (t_i, \delta_i) \right) = \frac{\mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \mathbb{P} \left( \boldsymbol{Z}_i = \boldsymbol{x}_j \right)}{\sum_{j=1}^{n_B} \mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \mathbb{P} \left( \boldsymbol{Z}_i = \boldsymbol{x}_j \right)},$$

with

$$\begin{aligned}
\mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) &= f(t_i \mid \boldsymbol{x}_j)^{\delta_i} S(t_i \mid \boldsymbol{x}_j)^{1-\delta_i} \\
&= \left[ \lambda_0(t_i) \exp \left( \boldsymbol{\beta}^\top \boldsymbol{x}_j \right) \right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j),
\end{aligned}$$

and

$$\begin{aligned} S(t_i \mid \boldsymbol{x}_j) &= \exp\left(-\int_0^{t_i} \lambda_0(s)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\, ds\right) \\ &= \exp\left(-\Lambda_0(t_i)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)\right) \quad \text{where} \quad \Lambda_0(t_i) = \int_0^{t_i} \lambda_0(s)ds. \end{aligned}$$

Replace the above expressions in the conditional expectation of the log-likelihood, and let

$$G_{ij}(\boldsymbol{\beta}) = \mathbb{P}\left(\boldsymbol{Z}_i = \boldsymbol{x}_j \mid (t_i, \delta_i)\right) = \frac{q_{ij}\left[\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j)}{\sum_{j=1}^{n_B} q_{ij}\left[\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j)}.$$

Then we obtain:

$$\begin{aligned} \mathbb{E}\left(\log\left(L(\boldsymbol{\beta}) \mid (T,\delta)\right)\right) &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \log\left[q_{ij}\left[\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j)\right] \times G_{ij}(\boldsymbol{\beta}), \\ &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \left[\log(q_{ij}) + \delta_i \log(\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)) + \log(S(t_i \mid \boldsymbol{x}_j))\right] \times G_{ij}(\boldsymbol{\beta}), \end{aligned}$$

# 5 Simulation studies

## 5.1 Data generation

We adapt the data generating process developped in Thanh et al (2022, 2023). We consider two databases $A$ and $B$ containing $K$ matching variables. We first generate the observations in database $B$ (which is of size $n_B$), and then, we extract a random subset of size $n_A$ of $B$, to obtain the database $A$. In database $B$, each $Y_j^k$ is simulated from a Bernoulli distribution with probability $p^k$, for $j = 1, \ldots, n_B$ and $k = 1, \ldots, K$ (here, we let $p^k = 0.2$ for every $k$). Since there are only binary matching variables, the record linkage methods require a large number of matching variables to get good performances. We choose $K = 40$. We consider $P = 2$ covariates, which are simulated as follows: $X_j^1 \simeq \mathcal{N}(0,1)$ and $X_j^2 \simeq \mathcal{B}(0.8)$. The true survival times $\tilde{T}_j$ are simulated as:

$$\tilde{T}_j = -\frac{\log(U_j)}{\lambda\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)},$$

where $U_j$ follow a standard uniform distribution. We set $\beta = (0.5, -0.5)^\top$ and $\lambda = 1$. Then, we obtain $T = \min(\tilde{T}, C)$ and $\delta = 1_{\tilde{T} \leq C}$ by using a fixed censoring time $C$, chosen to yield an approximate censoring rate equal to 0.25.
Finally, a random subset of $B$ is selected to produce the database $A$. We observe $(T, \delta)$ in database $A$ and only $\boldsymbol{X}$ in database $B$.

To account for possible errors in the matching variables, the $Y_i^k$ in $A$ (that is, for $i = 1, \ldots, n_A$) are obtained from the $Y_i^k$ in $B$ as:

$$Y_i^k = \begin{cases} Y_i^k & \text{with probability } 1 - e^k \\ 1 - Y_i^k & \text{with probability } e^k \end{cases} \qquad \text{for } k = 1, \ldots, K.$$

We choose $e^k = 0.04$. Let $\boldsymbol{Q} = (q_{ij})_{1 \leq i \leq n_A, 1 \leq j \leq n_B} \in \mathbb{M}_{(n_A, n_B)}$ , be a matching probability matrix where $q_{ij}$ is given by (2).

## 5.2 Methods

We consider two scenarios. We first assume that the $\boldsymbol{Q}$ matrix resulting from the record linkage process is known (here, we choose a set a random values for the $q_{ij}$). In a second step, we estimate $\boldsymbol{Q}$ for the data generated in section 5.1 by the record linkage method developed by Vo et al. (2023). For each of these scenarios, we compare the values and properties of the different estimators of $\boldsymbol{\beta}$ obtained with the methods 1, 2 and 3 proposed in section 4.2.

# Bibliographie


Fellegi, I. and Sunter, A. (1969), A theory for record linkage, *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Thanh, H., Chauvet, G., Happe, A., Oger, E., Paquelet, S., Garès, V. (2022), Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system, *Journal of Computational Statistics and data Analysis*, 179, n°107656 .

Cox, D. (1972), Regression models and life-tables, *Journal of the royal statistical society*, Series B (Methodological),ISSN 00359246, 34, pp. 187-220.

F.Santos. (2015), L'algorithme EM: une courte présentation, CNRS, UMR 5199 PACEA.

Meng, X. and Rubin, D. ( 1993), Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, pp. 267-278.

Hof, M., Ravelli, A. and Zwinderman, A. (2017), A probabilistic record linkage model for survival data, *Journal of the american statistical association*, 112(520), pp. 1504-1515.

Vo, TH., Garès, V., Zhang LC, Happe, A., Oger, E., Paquelet, S., and Chauvet, G. (2023), Cox regression with linked data, *Statistics in medicine* 43(2), pp. 296-314.

Ying, H. and Lahiri, P. (2019), Statistical analysis with linked data, *International statistical review* , 87( S1), pp S139-S157.

Lahiri, P. and Larsen, D. (2005), Regression analysis with linked data, *Journal of the American statistical association* , 100(469), pp. 222-230.

Danhyang, L., Li-Chun, Z. and Jea, K. (2022), Maximum entropy classification for record linkage, *Survey methodology*, 48(1), pp. 1-23.

Andersen, P., and Gill, R., (1982), Cox's regression model for counting processes: A large sample study, *The annals of statistics*, 10(4), pp 1100-1120.

# Variabilités intra et inter-visites de la pression artérielle et risque de démence : un modèle conjoint avec variance résiduelle individuelle

Léonie Courcoul [1,a] & Antoine Barbieri [2,a] & Christophe Tzourio [3,a] & Hélène Jacqmin-Gadda [4,a]

[a] *Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France*
[1] *leonie.courcoul@u-bordeaux.fr*
[2] *antoine.barbieri@u-bordeaux.fr*
[3] *christophe.tzourio@u-bordeaux.fr*
[4] *helene.jacqmin-gadda@u-bordeaux.fr*

**Résumé.** La démence dans le monde touche aujourd'hui près de 50 millions de personnes et ce nombre ne fait qu'augmenter. Etant donné l'absence de traitement, le levier d'action majeur reste la prévention sur les facteurs de risques modifiables tels que les facteurs cardio-vasculaires. Dans cette optique, il est aujourd'hui reconnu que l'hypertension artérielle est un facteur de risque de démence. Un nombre croissant d'études suggère que la variabilité de la pression artérielle pourrait également être un facteur de risque de démence. Cependant, ces études souffrent de faiblesses méthodologiques importantes et ne permettent pas de distinguer la variabilité à long-terme de celle à court-terme. L'objectif de ce travail est de développer un modèle conjoint avec une variance résiduelle individuelle flexible pour les mesures répétées d'un marqueur longitudinal et le risque d'événements compétitifs afin d'étudier l'association entre la variabilité de la pression artérielle et le risque de démence en tenant compte du risque compétitif de décès. La variance résiduelle d'un modèle mixte permet de distinguer la variabilité inter-visites de la variabilité intra-visite. Un effet aléatoire spécifique au sujet est inclu dans ces deux variabilités. Les risques peuvent dépendre simultanément de la valeur et de la pente courantes du marqueur ainsi que de chacune des deux composantes de la variance résiduelle. Le modèle a été estimé sur les données de la cohorte des trois-cités (3C) pour étudier l'impact des variabilités à court et long-terme de la pression artérielle sur le risque de démence et de décès.

**Mots-clés.** Modèle conjoint, variabilité intra et inter-visites, pression artérielle, démence.

**Abstract.** Dementia currently affects about 50 million people wordwilde, and this number is rising. Given that there is still no cure, the major level for action remains prevention of modifiable risk factors such as cardiovascular factors. With this in mind, it is now recognized that a high level of blood pressure is a risk factor for dementia. A growing number of studies suggest that blood pressure variability may also be a risk factor for dementia. However, these studies suffer from significant methodological weaknesses

and fail to distinguish long-term from short-term variability. The aim of this work was to develop a joint model with flexible individual residual variance for repeated measurements of a longitudinal marker and the risk of competing events in order to study the association between blood pressure variability and dementia risk, taking into account the competing risk of death. The residual variance of the mixed model distinguishes inter-visit variability from intra-visit variability. A subject-specific random effect is included in both variabilities. Risks can depend simultaneously on the current value and slope of the marker, as well as on each of the two components of the residual variance. The model was estimated on data from the Three-Cities (3C) cohort to study the impact of short- and long-term blood pressure variability on the risk of dementia and death.

**Keywords.** Joint model, intra- and inter-visit variability, blood pressure, dementia.

# 1 Introduction

La démence touche plus de 50 millions de personnes dans le monde et le nombre de cas continue d'augmenter avec l'espérance de vie, et sans traitement la prévention sur les risques vasculaires modifiables demeure le levier d'action principal. L'hypertension artérielle représente un facteur de risque bien établi pour le risque de démence. Par ailleurs, un nombre croissant d'études suggère que la variabilité de la pression artérielle pourrait également être un facteur de risque de ces événements, indépendamment du niveau de la pression artérielle [1]. Ces études s'intéressent majoritairement à la variabilité de la pression artérielle à long-terme, c'est-à-dire, celle calculée sur des mesures répétées au cours de plusieurs années. En comparaison, peu d'études s'intéressent à la variabilité de la pression artérielle à court ou moyen terme [2]. De plus, la plupart des études ne permettent pas de prendre en compte rigoureusement ces différents types de variabilités et souffrent de faiblesses méthodologiques. En effet, la variabilité de la pression artérielle est souvent calculée comme l'écart-type des mesures de la pression artérielle puis incluse comme un facteur de risque dépendant du temps dans un modèle de Cox. Cette approche est source de biais puisqu'elle néglige l'erreur de mesure sur l'écart-type de la pression artérielle et nécessite l'imputation de l'écart-type à tous les temps d'événement. De plus, la pression artérielle et son écart-type sont des variables endogènes, pour lesquelles le modèle de Cox n'est pas adapté [3,4,5].

Pour éviter ces biais un modèle conjoint combinant un modèle mixte incluant un effet aléatoire spécifique au sujet pour la variance résiduelle et un modèle à risques proportionnels a été proposé [6,7,8]. Cependant, avec ce modèle une seule mesure de la pression artérielle est prise en compte à chaque visite alors que dans la plupart des études, la pression artérielle est mesurée deux fois ou plus à chaque visite. Sur le plan clinique, il

serait intéressant d'évaluer si la variabilité à court terme (intra-visite) est prédictrice de démence car sa mesure est très aisée.

L'objectif de ce travail était donc d'étendre ce modèle conjoint afin de considérer la variance résiduelle du modèle mixte comme la somme de deux composantes individuelles : la variabilité intra- et inter-visites. Les risques d'événements peuvent ainsi dépendre simultanément de la valeur courante, de la pente et des variabilités intra- et inter-visites du marqueur. Les performances de la procédure d'estimation ont été vérifiées via une étude de simulation. Enfin, le modèle a été appliqué aux données de la cohorte des trois-cités [9] pour étudier rigoureusement l'impact des variabilités intra-visite et inter-visites de la pression artérielle sur le risque de survenue de la démence, tout en prenant en compte le risque compétitif de décès.

## 2 Le modèle

Considérons un échantillon de $N$ individus. Pour chaque individu $i \in \{1, ..., N\}$, $Y_{ijl}$ est la valeur du marqueur longitudinal pour la mesure $l = 1, \ldots, n_{ij}$, à la visite $j = 1, \ldots, n_i$. La visite $j$ a lieu au temps $t_{ij}$. À chaque visite $j$, l'individu $i$ est susceptible d'avoir $n_{ij}$ mesures du marqueur longitudinal. $Y_i$ est un vecteur de dimension $\sum_j n_{ij}$ contenant l'ensemble des mesures du marqueur pour l'individu $i$. Supposons qu'il existe deux événements compétitifs et une censure à droite. On note $T_{i1}^*$ et $T_{i2}^*$ les vrais temps pour les deux types d'événements du sujet $i$ et $C_i$ le temps de censure. Soit $T_i$ le temps d'événement observé défini par $T_i = \min(T_{i1}^*, T_{i2}^*, C_i)$. Soit $\delta_i \in \{0, 1, 2\}$ l'indicateur d'événement, avec $\delta_i = k$ si l'événement compétitif $k$ est observé et $\delta_i = 0$ en cas de censure.

Le modèle conjoint à effets aléatoires partagés est défini par un sous-modèle linéaire mixte et un sous-modèle de survie pour chaque événement. Le sous-modèle linéaire mixte s'écrit, pour l'individu $i$, à sa visite $j$ et pour sa mesure $l$ de la manière suivante :

$$\begin{cases} Y_{ijl} = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} + \nu_{ijl} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} + \nu_{ijl}, \\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2) \quad \text{with} \quad \log(\sigma_i) = \mu_\sigma + \tau_{\sigma i}, \\ \nu_{ijl} \sim \mathcal{N}(0, \kappa_i^2) \quad \text{with} \quad \log(\kappa_i) = \mu_\kappa + \tau_{\kappa i}, \end{cases} \tag{1}$$

avec $X_{ij}^\top$ et $Z_{ij}^\top$ deux vecteurs de covariables au temps $t_{ij}$. Ils sont respectivement associés au vecteur des effets fixes $\beta$ et au vecteur des effets aléatoires $b_i$. Dans ce modèle, l'erreur résiduelle correspond à la somme de deux composantes : l'erreur spécifique à la visite, $\epsilon_{ij}$, et l'erreur spécifique à la mesure $\nu_{ijl}$. Ces deux erreurs sont spécifiques au sujet via l'inclusion d'un effet aléatoire individuel pour chacune d'entre elles ($\tau_{\sigma_i}$ et $\tau_{\kappa_i}$ respectivement). Les effets aléatoires sont supposés corrélés entre eux :

$$\begin{pmatrix} b_i \\ \tau_{\sigma i} \\ \tau_{\kappa i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_b & \Sigma_{\tau_\sigma b} & \Sigma_{\tau_\kappa b} \\ \Sigma_{\tau_\sigma b}^\top & \Sigma_{\tau_\sigma} & \Sigma_{\tau_\sigma \tau_\kappa} \\ \Sigma_{\tau_\kappa b}^\top & \Sigma_{\tau_\sigma \tau_\kappa}^\top & \Sigma_{\tau_\kappa} \end{pmatrix}\right)$$

On définit le sous-modèle de survie pour l'événement $k \in \{1, 2\}$ par :

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp\left(W_i^\top \gamma_k + \alpha_{1k}\tilde{Y}_i(t) + \alpha_{2k}\tilde{Y}_i'(t) + \alpha_{\sigma k}\sigma_i + +\alpha_{\kappa k}\kappa_i\right), \tag{2}$$

avec $\lambda_{0k}(t)$ la fonction de risque de base, $W_i$ un vecteur de covariables avec les coefficients de régression correspondants $\gamma_k$, et $\alpha_{1k}$, $\alpha_{2k}$, $\alpha_{\sigma k}$ et $\alpha_{\kappa k}$ les coefficients de régression associés à l'effet de la valeur courante, de la pente courante et de chacune des deux composantes de la variabilité résiduelle, sur le risque de l'événement $k$.

Le risque de base peut être de type exponentiel ou Weibull, ou, pour plus de flexibilité, il peut être défini sur une base de B-splines par :

$$\log(\lambda_{0k}(t)) = \exp\left(\sum_{q=1}^{Q} \eta_{qk} B_q(t, \nu_k)\right) \tag{3}$$

où $B_q(t, \nu_k)$ dénote la q-ème fonction de base de B-splines avec le vecteur de noeuds $\nu_k$.

# 3 Méthode d'estimation

Notons $\theta$ l'ensemble des paramètres à estimer, incluant donc les paramètres de la décomposée de Cholesky de la matrice de covariance des effets aléatoires, $\beta$, $\mu^\top = (\mu_\sigma, \mu_\kappa)$, $\alpha^\top = (\alpha_{11}, \alpha_{21}, \alpha_{\sigma 1}, \alpha_{\kappa 1}, \alpha_{12},\ \alpha_{22}, \alpha_{\sigma 2}, \alpha_{\kappa 2})$, $\gamma^\top = (\gamma_1, \gamma_2)$ et les paramètres des deux fonctions de risque à baseline. L'estimation des paramètres se fait sous le paradigme fréquentiste par maximisation de la vraisemblance. La contribution de l'individu $i$ à la vraisemblance marginale du modèle avec variance résiduelle hétérogène, s'écrit :

$$\begin{aligned} \mathcal{L}_i(\theta; Y_i, T_i, \delta_i) &= \int p(Y_i, T_i, \delta_i | b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta) f(b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta) db_i d\tau_{\sigma i} d\tau_{\kappa i} \\ &= \int f(Y_i | b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta) \exp\left(-\sum_{k=1}^{2} \Lambda_{ik}(T_i | b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)\right) \times \\ &\quad \prod_{k=1}^{2} \lambda_{ik}(T_i | b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)^{1_{\delta_i = k}} f(b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta) db_i d\tau_{\sigma i} d\tau_{\kappa i}, \end{aligned}$$

avec :

- $f(b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)$ une densité multivariée Gaussienne

- $f(Y_i|b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta) = \prod_{j=1}^{n_i} f(Y_{ij}|b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)$ où $f(Y_{ij}|b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)$ est une densité multivariée Gaussienne de la matrice de covariance diagonale et symétrique suivante:
$$\begin{vmatrix} \sigma_i^2 + \kappa_i^2 & \sigma_i^2 & \dots & \dots & \sigma_i^2 & \sigma_i^2 \\ \sigma_i^2 & \ddots & \ddots & \ddots & \ddots & \sigma_i^2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \sigma_i^2 & \ddots & \ddots & \ddots & \sigma_i^2 + \kappa_i^2 & \sigma_i^2 \\ \sigma_i^2 & \sigma_i^2 & \dots & \dots & \sigma_i^2 & \sigma_i^2 + \kappa_i^2 \end{vmatrix}$$
- pour $k \in \{1, 2\}$, $\Lambda_{ik}(T_i|b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)$ la fonction de risque cumulée

L'intégrale sur les effets aléatoires n'ayant pas de solution analytique, elle est calculée numériquement par la méthode de quasi-Monte-Carlo. L'intégrale pour les risques cumulés est calculée par Gauss-Kronrod. La maximisation de la fonction de vraisemblance se fait par l'algorithme Marquardt-Levenberg, variante robuste de l'algorithme de Newton-Raphson, grâce à la fonction `marLevAlg` du package éponyme [10]. La procédure d'estimation a été validée par simulations.

# 4 Application

Le modèle a été appliqué à la cohorte des trois-cité (3C). Cette cohorte prospective française en population générale avait pour objectif principal l'étude de l'association entre les facteurs vasculaires et le risque de démence. Les 9031 individus de cette cohorte devaient être inscrits sur les listes électorales de Bordeaux, Dijon ou Montpellier, être âgés de 65 ans ou plus entre mars 1999 et mars 2001 et ne pas être institutionnalisés. Diverses mesures ont été collectées au cours du suivi, à 2, 4, 7, 10, 12, 14 et 17 ans après l'inclusion, comme la pression artérielle qui a été mesurée au minimum deux fois à chacune de ces visites. L'événement d'intérêt considéré est le diagnostic de démence et les décès dus à d'autres causes ont été considérés comme des événements concurrents. Le diagnostic de démence a été évalué à chaque visite et le temps de démence est défini comme le milieu de l'intervalle entre la dernière visite saine et la visite de diagnostic.

# Bibliographie

[1] Ma, Y., Tully, P.J., Hofman A., and Tzourio, C. (2020). Blood Pressure Variability and Dementia: A State-of-the-Art Review. *American Journal of Hypertension*, 33(12).

[2] Stevens, S.L., Wood, S., Koshiaris, C., Law, K., Glasziou, P., Stevens, R.J. and McManus, R.J. (2016). Blood pressure variability and cardiovascular disease: systematic

review and meta-analysis, *BMJ*.

[3] de Courson, H., Leffondre, K., Tzourio, C. (2018). Blood pressure variability and risk od cardiovascular event : is it appropriate to use the future for predicting the present ? *European heart journal*, 39, 4220-4220.

[4] Prentice, RL. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69, 331-342.

[5] Kalbeisch, J. and Prentice, R. (2002). The Statistical Analysis of Failure Time Data, 2nd edition. *John Wiley Sons*.

[6] Barrett, JK., Huille, R., Parker, R., Yano, Y., Griswold, M. (2019). Estimating the association between blood pressure variability and cardiovascular disease : An application using the aric study. *Statistics in medicine*, 38, 1855-1868.

[7] Gao F, Miller JP, Xiong C et al (2011). A joint-modeling approach to assess the impact of biomarker variabilityon the risk of developing clinical outcome. *Statistical Methods Applications* , 20(1): 83–100.

[8] Courcoul L., Tzourio C., Woodward M., Barbieri A., Jaqmin-Gadda H. (2023). A location-scale joint model for studying the link between the time-dependent subject-specific variability of blood pressure and competing events. arXiv:2306.16785.

[9] 3C Study Group (2003). Vascular factors and risk of dementia : design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, 22(6) :316–325.

[10] Philipps, V., Heijblum, B., Prague, M., Commenges, D., Proust-Lima, C. (2021). Robust and efficient optimization using a Marquardt-Levenberg algorithm with R package marqlevalg. *The R Journal*, 13.

# Construction de récompenses par apprentissage par préférences pour les modèles d'apprentissage par renforcement appliqués aux stratégies de traitements adaptatifs

Sophia Yazzourh[1] , Nicolas Savy [1], Philippe Saint-Pierre [1] et Michael Kosorok[2]

[1] *Institut de Mathématiques de Toulouse; UMR5219 - Université de Toulouse ; CNRS - UPS IMT,*
*F-31062 Toulouse Cedex 9, France*
[2] *Department of Biostatistics, University of North Carolina at Chapel Hill,*
*Chapel Hill, NC 27599, U.S.A.*

**Résumé.** Délivrer des traitements personnalisés à chaque étape des maladies chroniques est un objectif clé de la médecine de précision qui est formalisé par les "Dynamic Treatment Regimes". Ce cadre adapte les stratégies de traitement en se basant sur des règles de décision apprises à partir d'essais cliniques pour améliorer l'efficacité du traitement. L'utilisation de l'apprentissage par renforcement aide à déterminer ces règles en se basant sur leurs données individuelles et de leurs historiques médicaux. L'apprentissage de la stratégie de traitement repose sur des évaluations quantitatives du système appelées récompenses. Traditionnellement, ces récompenses sont déterminées par des experts qui sélectionnent une variable d'intérêt, mais qui peut être trop restrictive pour l'apprentissage de la règle de décision. Nous proposons une approche automatique et généralisée pour construire les récompenses, en utilisant l'apprentissage par préférences ou "Preference Learning".

**Mots-clés.** Apprentissage par renforcement; Stratégies de traitements adaptatifs; Médecine de précision; Apprentissage par préférences; Récompenses;

**Abstract.** Delivering personalized treatments at each stage of chronic diseases is a key goal of precision medicine, formalized by "Dynamic Treatment Regimes." This framework adjusts treatment strategies based on decision rules learned from clinical trials to enhance treatment effectiveness. Reinforcement learning helps determine these rules by using individual patient data and medical histories. Learning treatment strategy relies on quantitative system evaluations called rewards. Traditionally, experts select these rewards based on a variable of interest, which may be too restrictive for learning the decision rule. We propose an automatic and generalized approach to construct rewards using preference learning.

**Keywords.** Reinforcement Learning; Dynamic Treatment Regimes; Precision Medicine; Preference Learning; Rewards;

# 1 Introduction

La médecine moderne cherche à améliorer sa capacité à traiter de manière personnalisée chaque patient. C'est notamment la médecine de précision [Kosorok and Laber, 2019] qui initie une réflexion profonde sur cette question. Son objectif est d'améliorer la qualité des soins de santé en ajustant l'approche médicale selon l'état de santé spécifique et changeant de chaque patient. L'hétérogénéité des caractéristiques et des réactions parmi les populations de patients exigent des approches de traitement variées. Initialement, des modèles statistiques ont été implémentés pour répondre à cette problématique. Avec l'avènement du stockage de données et de la puissance de calcul, les méthodes d'apprentissage automatique ont également commencé à être appliquées comme le montre Yu et al. [2021] et Coronato et al. [2020].

Dans ce contexte, la médecine moderne s'intéresse de plus en plus à l'adaptation des traitements aux données individuelles des patients. La médecine de précision vise à améliorer la santé en mettant au coeur de la décision les informations spécifiques à chaque patient. Cette approche qui cherche à recommander des traitements personnalisés est appelée "Dynamic Treamtent Regimes" (DTR) et est explicitée dans Kosorok and Laber [2019] et Chakraborty and Murphy [2014]. Les DTR se basent sur l'historique médical et les réponses des patients aux traitements précédents.

Au fil des décennies, l'apprentissage automatique est devenue une solution incontournable pour résoudre des problèmes complexes à grande échelle. Dans le domaine du support à la décision, particulier pour les scénarios séquentiels, l'apprentissage par renforcement ou "Reinforcement Learning" (RL) [Sutton and Barto, 2018] se révèle être une approche particulièrement efficace. Ces méthodes s'adaptent aux conditions changeantes et optimisent les décisions sur plusieurs étapes. Elles répondent à des questions de processus de prise de décisions dynamiques. L'idée principale est de trouver une règle de décision, appelée "policy", visant à optimiser un objectif à long terme, en prenant des décisions successives pour maximiser le bénéfice global. Cet objectif d'optimisation est basé sur des récompenses qui sont des indications quantitatives sur l'état du patient. Leur construction ou formulation est alors cruciale dans l'apprentissage de la prise de décision.

De manière générale, les récompenses sont conçues par un expert du système qui propose de l'évaluer au travers d'un score. Dans le contexte des essais cliniques visant à aider les personnes obèses à perdre du poids, une récompense pourrait consister à mesurer leur indice de masse corporelle (IMC) [Linn et al., 2015]. Dans le contexte des soins critiques, un autre exemple serait d'évaluer les traitements en fonction du taux de survie ou de mortalité des patients [Roggeveen et al., 2021]. Certaines récompenses peuvent être conçus de manière plus subtile en faisant des compromis et des combinaisons de variables. Dans le cadre d'une simulation de cancer [Zhao et al., 2009], les récompenses sont évaluées en prenant en compte la taille de la tumeur, de la toxicité du traitement, du bien être du patient et du taux de survie.

Construire de manière manuelle une fonction récompense peut impliquer des choix arbitraires, voire très spécifiques au contexte mais également mener à des objectifs d'apprentissages trop restrictifs. Dans la simulation de [Zhao et al., 2009], lorsque le décès d'un patient survient, un score de -60 est arbitrairement attribué à cet événement. Une des manière de

généraliser et de construire automatiquement des récompenses est d'utiliser l'apprentissage par préférences ou "Preference Learning" (PL) [Fürnkranz et al., 2012]. L'approche repose sur l'expertise médicale, où le médecin exprime ses préférences quant aux trajectoires ou aux suivis médicaux des patients. Ces informations de comparaison deux à deux seront ensuite utilisées au travers d'un modèle probabiliste comme celui de Bradley-Terry [Akrour et al., 2012] pour construire par maximum de vraisemblance les récompenses.

Dans la suite nous proposons un aperçu de l'application du RL à l'optimisation des séquences de traitement et une amélioration de la construction des récompenses par apprentissage par préférences.

# 2 Apprentissage par renforcement

## 2.1 Processus de Décision

Ce contexte de modélisation est celui des processus de décision ou "Decision Process" (DP) qui sont le cadre initial des DTR. Un DP est un système dynamique au cours du temps $t \in \mathbb{T}$ qui navigue dans un espace d'états $\mathbb{S}$ contenant les covariables décrivant l'état du patient. Les possibles actions sont contenues dans l'espace $\mathbb{A}$ et représentent les traitements et leurs dosages associés. Le sous-ensemble mesurable non vides de $\mathbb{A}$, notée $\{\mathbb{A}(s)|s \in \mathbb{S}\}$, contient les actions réalisables qui peuvent être prises lorsque le système se trouve dans un état spécifique $s \in \mathbb{S}$. En d'autres termes tous les traitements ne sont pas admissibles ou disponibles pour un patient à une étape donnée. Le formalisme mathématiques est donné par la définition suivante :

**Definition 2.1** (Processus de décision ou "Decision Process" (DP))**.** Un processus de décision $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$ sur $\mathbb{T}$ contient:

- une famille $S$ de variables aléatoires à valeurs dans $\mathbb{S}$ : $\{S_t, t \in \mathbb{T}\}$, où $\mathbb{S}$ est appelé espace d'états.
- une famille $A$ de variables aléatoires à valeurs dans $\mathbb{A}$ : $\{A_t, t \in \mathbb{T}\}$, où $\mathbb{A}$ est appelé espace d'actions.
- une famille $\{\mathbb{A}(s)|s \in \mathbb{S}\}$ des sous-ensembles mesurables non vides de $\mathbb{A}$, ensemble des actions réalisables lorsque le système se trouve dans l'état $s \in \mathbb{S}$. Cela requière à $\mathbb{K} = \{(s, a)|s \in \mathbb{S}, a \in \mathbb{A}(s)\}$ d'être un sous ensemble mesurable de $\mathbb{S} \times \mathbb{A}$.
- une distribution initiale $\nu$ sur $\mathbb{S}$.

**Definition 2.2** (Histoires admissibles)**.** Pour tout $n \in \mathbb{N}$, une histoire admissible au temps $n$ est un vecteur qui contient les états parcourus par le système ainsi que les actions prises jusqu'au temps $n$. L'ensemble des histoires admissibles au temps $n$ est noté:

$$\mathbb{H}_0 = \mathbb{S} \qquad \mathbb{H}_n = \mathbb{K}^{n-1} \times \mathbb{S} \tag{1}$$

Un élément $h_n \in \mathbb{H}_n$ s'écrit $(s_0, a_0, \ldots, s_{n-1}, a_{n-1}, s_n)$ où pour tout $0 \leq j \leq n-1,\ (s_j, a_j) \in \mathbb{K}$.

Les histoires admissibles $h_n$ observées sont les trajectoires de soins de différents patients et décrivent les informations des traitements et des covariables à chaque étape.

L'aspect principal à considérer dans le traitement du processus de décision est de déterminer la probabilité d'atteindre l'état $s_{n_1}$ au temps $n+1$ étant donné l'historique et les décisions jusqu'au temps $n$. On l'exprime comme suit :

$$\mathbb{P}_\nu \left[ S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right].$$

En pratique, le calcul de ces probabilités demande des ressources computationnelles significatives en raison de l'augmentation de la longueur du vecteur $h_n$, proportionnellement à $n$. Travailler directement avec une telle variable devient rapidement inabordable (généralement lorsque $n \geq 4$). De manière classique, cette difficulté est surmontée par l'hypothèse markovienne. Ainsi, le formalisme traditionnel du RL est souvent celui des processus de décision markovien.

## 2.2 Stratégie

Un des principaux concepts du RL est celui de stratégie, politique ou "policy". Elle correspond à la stratégie ou règle de décision médicale personnalisée que les DTR cherchent à déterminer à chaque étape d'intervention. Un seul traitement est administré par étape.

**Definition 2.3.** Une stratégie est une séquence $\pi = (\pi_n)_{n \in \mathbb{N}}$ de distributions conditionnelles de $\mathbb{A}$ sachant $\mathbb{H}_n$ définit, pour tout $\mathcal{A} \in \mathcal{B}(\mathbb{A})$ et pour tout $h_n \in \mathbb{H}_n$, par:

$$\pi_n(\mathcal{A}, h_n) = \mathbb{P}\left[ A_n \in \mathcal{A} \mid H_n = h_n \right],$$

satisfaisant pour tout $n \in \mathbb{N}$, pour tout $h_n \in \mathbb{H}_n$ :

$$\pi_n(\mathbb{A}(s_n), h_n) = 1,$$

et pour tout $n \in \mathbb{N}$, pour tout $h_n \in \mathbb{H}_n$ et pour tout $a_n \in \mathbb{A}(s_n)$

$$\pi_n(a_n, h_n) > 0.$$

La prise de décision implique la sélection d'une option en fonction des informations environnementales. Une politique représente un plan stratégique aligné sur un objectif spécifique définissant une séquence d'actions. Une stratégie $\pi$ suggère une action $a_n$ pour chaque état possible $s_n$ en tenant compte de l'historique $h_n$.

## 2.3 Récompense et optimalité

L'objectif qui guide la stratégie est formalisé au travers d'un critère d'optimalité appelé récompense ou "reward". Dans un context d'optimisation de traitement, les récompenses

sont calibrées pour répondre à un objectif médical. Elles peuvent être déterminées par l'expertise des médecins ou selon des objectifs précis.

**Definition 2.4.** Les récompenses $\{R_n, n \in \mathbb{N}\}$ sont les éléments d'une famille de variables aléatoires bornées dans $\mathbb{R}$.

Afin de pouvoir évaluer et comparer les stratégies, deux mesures quantitatives doivent êtres introduites :

**Definition 2.5** (Fonctions valeurs [Chakraborty and Murphy, 2014])**.** Étant donné un processus de décision $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$ sur $[0, \tau]$, $\{R_n, n \in \mathbb{N}\}$ une famille de récompenses et $\pi$ une stratégie :

- A l'étape $n$ la V-fonction ou "state-value function" est l'espérance de la somme des récompenses attendues depuis l'étape $n$ sachant l'histoire $h_n$:

$$V_n^\pi(h_n) = \mathbb{E}_\nu^\pi \left[ \sum_{j=n+1}^{\tau} R_j \mid H_n = h_n \right]. \tag{2}$$

- A l'étape $n$ la Q-fonction ou "action-value function" est l'espérance de la somme des récompenses attendues depuis l'étape $n$ sachant l'histoire $h_n$ et l'action prise $a_n$:

$$Q_n^\pi(h_n, a_n) = \mathbb{E}_\nu^\pi \left[ \sum_{j=n+1}^{\tau} R_j \mid H_n = h_n, A_n = a_n \right]. \tag{3}$$

L'objectif du RL est de déterminer les stratégies optimales notées $\pi^*$ qui sont les politiques qui maximisent la somme des récompenses. En d'autres mots, celles qui maximisent le gain à long terme. Les stratégies optimales peuvent alors être déterminées en évaluant les fonctions valeurs optimales.

**Theorem 2.1.** Les stratégies optimales $\pi*$ sont les stratégies qui maximisent les fonctions valeurs pour tout $n \in \mathbb{N}$, pour tout $h_n \in \mathbb{H}_n$ et $a_n \in \mathbb{A}$, tel que

$$V_n^{\pi^*}(h_n) = V_n^*(h_n) = \max_\pi \; V_n^\pi(h_n) \qquad \text{et} \qquad Q_n^{\pi^*}(h_n, a_n) = Q_n^*(h_n, a_n) = \max_\pi \; Q_n^\pi(h_n, a_n).$$

# 3 Construction de récompenses par apprentissage par préférences

## 3.1 Apprentissage par préférences

L'apprentissage par préférences ou "Preference Learning" (PL) [Fürnkranz et al., 2012] offre une alternative à la construction de récompenses. Dans le cadre de l'optimisation des séquences de traitement, les récompenses sont ajustées pour s'aligner sur les objectifs médicaux. Cependant, cette approche se limite parfois à la maximisation d'une seule mesure

quantitative, ce qui peut être trop restrictif pour l'apprentissage. Le PL propose de construire des récompenses $\{R_n, n \in \mathbb{N}\}$ en se basant sur l'avis d'un expert. L'approche consiste à solliciter un avis médical en demandant la préférence du médecin entre deux éléments spécifiques. Ensuite, un modèle probabiliste permets de construire les récompenses qui seront par la suite incorporées dans une méthode d'apprentissage par renforcement classique afin de déterminer les stratégies $\pi*$ qui les maximisent.

La préférence ou comparaison des éléments deux à deux peut porter sur des trajectoires de patients $h_n$, des stratégies de traitements $\pi$, des états $s_n$ ou des actions $a_n$. Dans notre modèle, nous nous intéressons aux préférences sur les histoires admissibles ou trajectoires de patients $h_n$. Une fois les préférences exprimées par l'expert, une relation d'ordre entre les éléments est établie. Puis le modèle probabiliste utilisé est celui de Bradley-Terry [Akrour et al., 2012], afin de comparer des trajectoires deux à deux, adapté ici aux DTR. Si l'on considère des préférences sur des trajectoires tel que $h^i \prec h^j$, notre objectif est de maximiser la probabilité suivante :

$$\mathbb{P}(h^i \prec h^j|\theta) = \frac{e^{\alpha(R_\theta(h^i)-R_\theta(h^j))}}{1+e^{\alpha(R_\theta(h^i)-R_\theta(h^j))}}.$$

On calcule par maximum de vraisemblance les récompenses paramétrées $R_\theta$ .

## 3.2 Application

On se place dans un contexte applicatif de simulation d'un cancer traité par chimiothérapie à 4 facteurs [Zhao et al., 2009] : (1) la croissance tumorale en l'absence de chimiothérapie ; (2) les résultats négatifs sur le bien-être des patients en réponse à la chimiothérapie ; (3) la capacité du médicament à tuer les cellules tumorales tout en augmentant la toxicité ; et (4) une interaction entre les cellules tumorales et le bien-être du patient. Pour chaque patient on a deux variables d'états $S_t = \{Y_t, X_t\}$ avec $Y$ la taille de la tumeur et $X$ la toxicité du traitement à chaque mois $t$ tel que $t = 0, 1, 2, 3$. Le traitement $A_t$ administré au mois $t$ est un dosage compris entre 0 et 1 avec un pas de 0,1. Ce modèle se base sur le système d'équations différentielles suivant :

$$\begin{aligned}\Delta Y_t &= [0,15 \times \max(X_t, X_0) - 1,2 \times (A_t - 0,5)] \times \mathbb{1}(Y_t > 0)\\ \Delta X_t &= 0,1 \times \max(Y_t, Y_0) + 1,2 \times (A_t - 0,5)\end{aligned}$$

En utilisant l'indicatrice $\mathbb{1}(Y_t > 0)$, le modèle attribue le statut de rémission totale à un patient lorsque la taille de sa tumeur est réduite à zéro, indiquant ainsi l'absence de récidive.

La possibilité de décès d'un patient pendant un traitement est représentée par un modèle de survie. Pour chaque intervalle de temps $(t-1, t]$, le taux de survie est défini comme une fonction de la taille de la tumeur et de la toxicité : $\lambda(t) = \exp(-4+Y_t+X_t)$. Dans ce modèle, la taille de la tumeur et la toxicité ont une influence tout aussi importante sur la survie du patient. La probabilité de décès du patient pendant l'intervalle de temps $(t-1, t]$ est :

$$\mathbb{P}_{décès} = 1 - \exp(-\int_{t-1}^{t} \lambda(x)dx)$$

Dans le modèle proposé par Zhao et al. [2009], les récompenses sont accordées en fonction du statut de survie, du bien-être du patient et de la taille de la tumeur. On notera que si le patient décède, la "récompense" était de -60. Ce qui corresponds à un choix arbitraire. Notre projet est de construire ces récompenses en se basant sur les préférences d'un expert.

On se base sur le modèle de Fürnkranz et al. [2012], qui propose d'exprimer des préférences sur les trajectoires de la manière suivante :

- Si le patient $h_j$ survit plus longtemps que le patient $h_i$ : $h^i \prec h^j$
- Deux patients décédés au même temps $t$ ne peuvent pas être comparées
- Si les patients survivent dans les deux trajectoires, on note $T^i$ et $T^j$ les toxicités maximales sur la durée de la trajectoire pour les patients $i$ et $j$ respectivement et $D^i$ et $D^j$ la taille de leurs tumeurs à la fin. Les trajectoires sont alors comparée par la dominance de Pareto suivante :
$$h^i \prec h^j \Leftrightarrow (T^j \leq T^i) \text{ et } (D^j \leq D^i)$$

# 4 Conclusion

La conception directe des récompenses nécessite une compréhension approfondie du processus de traitement médical. En revanche, l'apprentissage par préférences offre une alternative à leur création, permettant d'intégrer davantage d'informations d'évaluation que simplement une variable quantitative.

Les récompenses obtenues grâce à notre modèle probabiliste de comparaison de trajectoires reposent sur un modèle de régression. Ce modèle, facilement interprétable, nous permettra de mettre en évidence la sélection pondérée des covariables de l'espace d'état. Nous pourrons alors comparer les récompenses obtenues par l'apprentissage par préférences avec celles construites manuellement. Finalement, nous étudierons et comparerons les deux règles de décision respectives obtenues par apprentissage par renforcement.

# References


Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020.

Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Kristin A Linn, Eric B Laber, and Leonard A Stefanski. iqlearn: Interactive q-learning in r. *Journal of statistical software*, 64(1), 2015.

Luca Roggeveen, Ali El Hassouni, Jonas Ahrendt, Tingjie Guo, Lucas Fleuren, Patrick Thoral, Armand RJ Girbes, Mark Hoogendoorn, and Paul WG Elbers. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. *Artificial Intelligence in Medicine*, 112:102003, 2021.

Yufan Zhao, Michael R Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315, 2009.

Johannes Fürnkranz, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89:123–156, 2012.

Riad Akrour, Marc Schoenauer, and Michèle Sebag. April: Active preference learning-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 116–131. Springer, 2012.

# Algorithmes statistiques pour la détection d'interactions médicamenteuses

Jules Bangard [1] & Étienne Birmelé [1]

[1] *Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS 7 rue René-Descartes, 67000 Strasbourg, France,* {***jbangard***, ***birmele***}*@unistra.fr*

**Résumé.** Une méthode statistique computationnelle est proposée pour la détection d'interactions médicamenteuses à risque, mettant en avant les enjeux de la surveillance post-commercialisation des médicaments. Cette étude met en œuvre un algorithme de Monte Carlo par Chaîne de Markov (MCMC) et un algorithme génétique pour identifier ces interactions à partir de données de pharmacovigilance. Une analyse des performances de l'algorithme MCMC, réalisée sur des données simulées, a montré des résultats très satisfaisants.

**Mots-clés.** Interactions Médicamenteuses, Algorithme MCMC, Algorithme Génétique, Optimisation Combinatoire, Pharmacovigilance

**Abstract.** A computational statistical method is proposed for detecting at-risk drug-drug interactions, highlighting the challenges of post-marketing drug surveillance. This study implements a Monte Carlo Markov Chain (MCMC) algorithm and a genetic algorithm to identify these interactions from pharmacovigilance data. An analysis of the MCMC algorithm's performance, conducted on simulated data, showed very satisfactory results.

**Keywords.** Drug-Drug-Interactions, MCMC Algorithm, Genetic Algorithm, Combinatorial Optimization, Computational Model, Pharmacovigilance

# 1 Introduction

## 1.1 Problème

### 1.1.1 Pharmacovigilance

Les phases d'essais cliniques, cruciales pour l'autorisation de mise sur le marché des médicaments, sont souvent limitées en taille et en diversité de profils médicaux (enfants, personnes immunodéprimées ...). Malgré leur importance, ces essais peuvent ne pas révéler certains effets secondaires qui se manifestent seulement après une utilisation prolongée. Cette lacune souligne la nécessité d'un deuxième rempart, la pharmacovigilance, qui surveille les risques d'effets indésirables post-commercialisation (Wikipédia (2024)).

La pharmacovigilance repose sur la collecte d'informations provenant des professionnels de santé et des industriels à travers des rapports individualisés appelés ICSR (**I**ndividual

**C**ase **S**afety **R**eport). Ces rapports contiennent des détails essentiels sur le patient, tels que son âge, ainsi que des informations sur les médicaments consommés et les effets rencontrés.

Grâce à l'émergence des différentes bases de données répertoriant ces ISCR, des méthodes d'exploitation de ces rapports ont émergé (Bate & Evans (2009)). Ces méthodes ont pour ambition la détection de "signaux", aiguillant ainsi les chercheurs en pharmacologie dans leur recherche de médicaments à risque.

### 1.1.2 Interactions médicamenteuses

Dans le domaine de la pharmacovigilance, l'attention s'est historiquement concentrée plus particulièrement sur l'identification d'effets secondaires résultant de la prise d'un unique médicament. Cependant, l'enjeu des interactions médicamenteuses est devenu un domaine de recherche de plus en plus étendu. Avec l'augmentation constante du nombre de médicaments disponibles, il devient impossible pour les pharmacologues d'examiner toutes les combinaisons possibles de manière exhaustive. Cette complexité est rendue préoccupante par certaines découvertes indiquant que divers traitements poly-médicamenteux offrent de meilleurs taux de rétablissement comparés aux traitements utilisant un unique médicament (Walkup et al. (2008)). Ainsi, il est essentiel de développer des méthodes capables d'évaluer le risque d'effets secondaires résultant de telles combinaisons. Pour cette raison, la méthode présentée a pour principal objectif la détection d'interactions médicamenteuses provoquant des effets secondaires chez les patients.

## 1.2 Données

### 1.2.1 Arbre des médicaments

Les médicaments sont organisés en arbre selon le système de classification Anatomique, Thérapeutique et Chimique (ATC) disposant de 5 niveaux de hiérarchie différents. Le plus haut niveau de hiérarchie est l'organe anatomique sur lequel agit le médicament, et le plus petit niveau la substance chimique classée. L'arbre des médicaments comptabilise un total de noeuds avoisinant les 5800. Dans la suite, les feuilles sont assimilées à un médicaments tandis que les autres noeuds à une famille de médicaments.

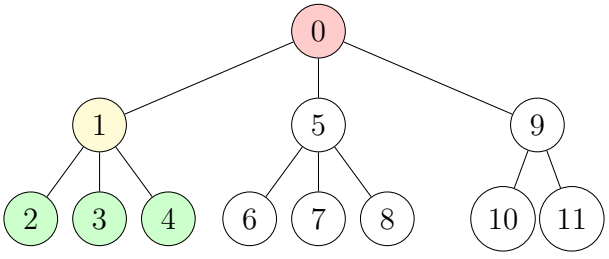



FIGURE 1 – Exemple simplifié d'arbre de médicaments

### 1.2.2 Cocktail de médicament

Pour représenter un cocktail de médicaments, l'arbre est numéroté d'après un algorithme de parcours en profondeur comme on peut le voir sur la Figure 1. Cette numérotation est utile par la suite pour encoder les cocktails. Un cocktail est une séquence de bits de taille $n$, où $n$ est le nombre de noeuds dans l'arbre. Un bit à l'indice $i$ vaut 1 si le médicament représenté par le noeud $i$ de l'arbre est pris par le patient et 0 sinon. On a par exemple, pour un patient prenant les médicaments 2 et 8 de l'arbre 1, la séquence $S$ suivante :

$$S = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Les bases de données de pharmacovigilance contenant les ICSR regroupent donc plusieurs séquences, toutes de taille $n$. Elles décrivent la prise de médicaments de chaque patient enregistré. Ces bases de données contiennent également les effets secondaires déclarés par les patients.

## 1.3 Caractérisation du risque associé à un cocktail

Ces données permettent d'associer un risque à une combinaison de médicaments. Il est possible de caractériser un tel risque à l'aide de différents scores. L'un des plus rependus de par sa facilité d'interprétation est le *Proportionnal Reporting Ratio* (PRR) proposé par (Evans et al. (2001)) et défini comme suit :

$$\frac{\mathbb{P}(AE|C)}{\mathbb{P}(AE|\neg C)}$$

où $AE$ est un effet secondaire et $C$ un cocktail de médicament.

Ce score est fréquemment utilisé en analyse de disproportionnalité (Bate et al. (1998); Evans, Waller, & Davis (2001); van Puijenbroek et al. (2002); Norén, Sundberg, Bate, & Edwards (2008)). Cette méthode utilise une table de contingence réalisée à l'aide d'une matrice regroupant les médicaments pris par chaque patient, ainsi qu'une matrice regroupant les effets secondaires subis par chaque patient encodée de la même manière que les cocktails de médicaments. Elle présente des avantages comme le temps de calcul qui est moindre mais également des inconvénients. Parmi eux, on peut citer les problèmes de masquages et de co-prescriptions (Maignen et al. (2014)).

De plus, certains scores ont été proposés pour généraliser ceux utilisés sur un unique cocktail comme le CRR ou le CSS qui sont des généralisations du PRR aux cocktails de taille supérieure à 1 (Noguchi et al. (2020)).

Une autre manière de caractériser le risque d'un cocktail est la suivante. Soient $n$ le nombre de patient prenant le cocktail $C$, $p$ la proportion de patient ayant l'effet secondaire $AE$ et $N$ le nombre total de patient dans le jeu de données. On définit $X$ comme la nombre de personnes ayant pris le cocktail $C$ et ayant subi l'effet secondaire $AE$. On prend comme score reflétant le risque

$$-\log(\mathbb{P}(X \geq x))$$

avec $X \sim \mathcal{H}(n, p, N)$ où $\mathcal{H}$ désigne la loi hypergeométrique.

# 2 Méthodes

L'identification de combinaisons de médicaments à haut risque est abordée via deux approches computationnelles. La première repose sur l'utilisation d'un algorithme de Monte-Carlo par chaînes de Markov (MCMC) pour l'exploration de l'espace des combinaisons de médicaments de taille $p$. Cette méthode permet d'estimer la distribution du risque associé à ces combinaisons. Ainsi il est possible de proposer une p-valeur empirique liée au score observé d'un cocktail. La seconde méthode, basée sur un algorithme génétique (Goldberg (2013)), vise à identifier de manière ciblée les combinaisons présentant un risque élevé, sans nécessiter une couverture exhaustive de l'espace des solutions.

## 2.1 Approximation du risque à travers les cocktails de médicaments

L'algorithme MCMC utilisé est l'algorithme de Metropolis-Hastings. Pour utiliser un tel algorithme, il faut définir un espace d'états $S = \{S_1, \ldots, S_p\}$. Il nécessite une mesure cible $f(S_i)$ calculable et des lois conditionnelles $q(.|S_i)$ sous lesquelles on sait simuler et grâce auxquelles il va pouvoir proposer de nouveaux états.

### Ensemble d'états

Un état est décrit par un cocktail de médicaments comportant $k$ médicaments, $k$ étant un hyperparamètre qui n'évolue pas au cours de l'algorithme.

Les états explorés peuvent contenir des noeuds internes à l'arbre (représentant donc des familles de médicaments). Cela permet la détection de signaux plus généraux. Par exemple, le paracétamol pourrait renvoyer un faible signal tandis que si on remonte dans l'arbre, les antalgiques pourraient peut-être représenter un signal plus général. Ainsi tous les patients prenant au moins un médicament de cette famille de médicaments seront considérés.

### Loi de proposition

Pour passer d'un cocktail à un autre, deux "mutations" différentes peuvent être effectuées, la mutation de type 1 et la mutation de type 2. Ces deux mutations sont complémentaires et exploitent la structure d'arbre des médicaments. Elles fonctionnent de la manière suivante.

**Mutation de type 1** La mutation de type 1 consiste en un mouvement totalement aléatoire dans l'espace des cocktails.

**Mutation de type 2** La mutation de type 2 consiste en un mouvement dit "local" vis-à-vis de la structure d'arbre des médicaments. En effet, lors de cette mutation on change un noeud de la séquence $S_i$ en l'un de ses noeuds voisins libres.

La Figure 2 représente un exemple d'une mutation de type 1 et 2. Pour la mutation de type 2, initialement la séquence contient les noeuds 2 et 3 en vert. On aperçoit, en orange à l'aide des arêtes orientées, les mouvements que la séquence peut effectuer. Un mouvement (une arête) est tiré uniformément parmi ceux disponibles. Dans notre

exemple, l'arête allant du noeud 2 vers la feuille 6 est choisie, ainsi, le noeud 2 est supprimé de la séquence et le 6 quant à lui est ajouté.

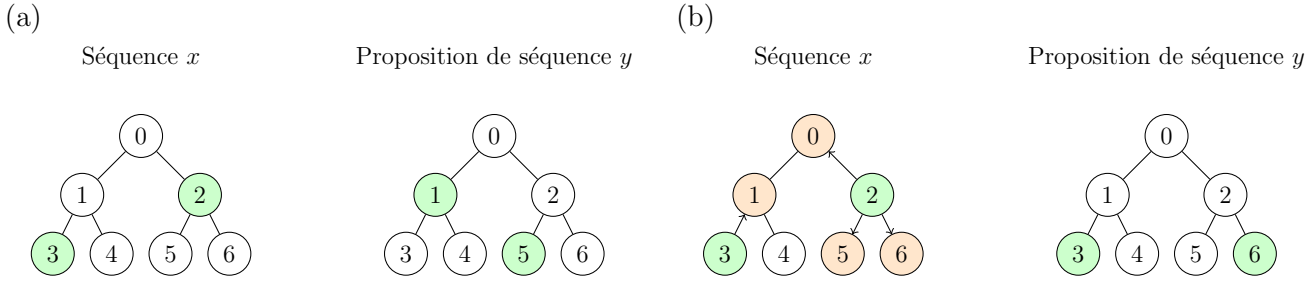



FIGURE 2 – (a) Exemple d'une mutation de type 1 (b) Exemple d'une mutation de type 2

La mutation de type 1 est proposée avec probabilité $p_I$ à chaque itération, $p_I$ étant un hyperparamètre. La mutation de type 2 est ainsi effectuée avec probabilité $p_{II} = 1 - p_I$.

### Évaluation de l'état

L'évaluation d'un cocktail de médicaments repose sur l'un des scores présenté précédemment, noté $H(S)$. La mesure cible $f$ choisie est alors la suivante :

$$f_T(S_i) = \frac{1}{Z(T)} \times e^{\frac{H(S_i)}{T}}$$

Où $Z(T) = \sum_S e^{\frac{H(S)}{T}}$. $T$ est un paramètre appelé température qui permet de moduler l'exploration de l'espace en acceptant plus aisément les cocktails de score modéré ($T$ élevé) ou, au contraire, en privilégiant fortement les combinaisons de médicaments de score élevé ($T$ faible).

La probabilité d'acceptation du cocktail $S_{i+1}$ à partir du cocktail $S_i$ est :

$$\min(1, \frac{f_T(S_{i+1})}{f_T(S_i)} \times \frac{q(S_i|S_{i+1})}{q(S_{i+1}|S_i)})$$

La théorie liée à l'algorithme de Metropolis-Hastings assure que la loi empirique des $f(S_i)$ pour la chaîne de cocktail ainsi construite converge vers la loi de $f(S)$. Une réalisation très longue d'une telle marche permet donc d'obtenir une loi approchée qui permet de déterminer une p-valeur empirique pour le score d'un cocktail d'intérêt.

## 2.2 Recherche des cocktails présentant le plus gros risque

L'algorithme génétique suit le modèle habituel de ce type d'algorithmes comme le montre la Figure 3. Il fait évoluer une population dans le but d'obtenir comme résultat une population performante au vu d'un critère d'évaluation arbitraire. Les étapes nécessaires pour cela sont :

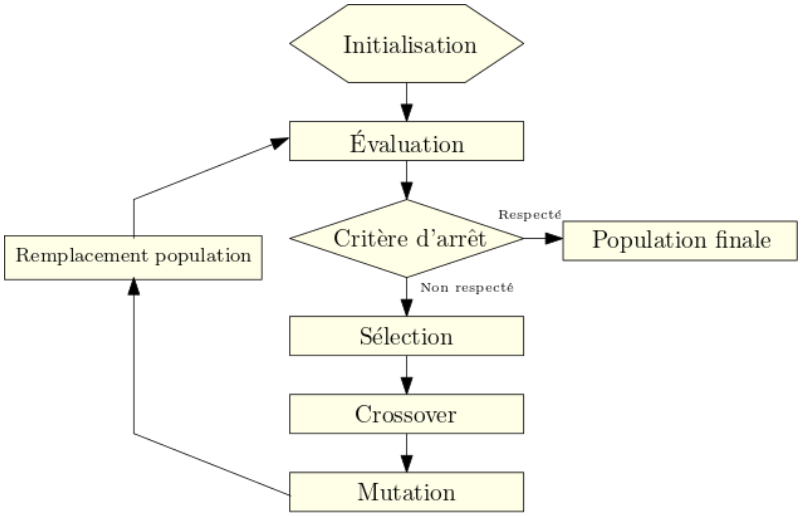



FIGURE 3 – Déroulement de l'algorithme génétique

**Initialisation** La population de l'algorithme génétique est un ensemble de $m$ cocktails de médicaments. Ces cocktails sont initialisés de manière aléatoire et peuvent être de tailles différentes contrairement à l'algorithme MCMC.

**Évaluation & Sélection** À chaque itération, la population passe à travers une phase d'évaluation et de sélection. L'évaluation calcule, pour chaque cocktail, son score hypergéométrique présenté en partie 1.3. Les scores des cocktails qui se ressemblent au sein de la population sont ensuite pénalisés. Le but de cette pénalisation est d'obtenir une population qui n'est pas trop homogène de sorte à obtenir plusieurs combinaisons médicamenteuses préoccupantes en sortie.

Les scores ainsi obtenus permettent d'effectuer des tournois consistant à tirer $k$ individus de la population et à conserver le meilleur des $k$ pour la phase de reproduction. De tels tournois sont effectués jusqu'à obtenir le nombre d'individus désiré pour la phase de reproduction.

**Modification & Remplacement de population** L'évolution de la population vers une population performante vis à vis du critère d'évaluation se fait en deux temps.

Dans un premier temps, une opération appelée crossover permet à deux séquences d'échanger de l'information. Dans le cas présent, le crossover consiste en l'échange de sous-arbres entre deux séquences de la manière suivante :

— Un noeud **interne** v de l'arbre est sélectionné aléatoirement.

— Les noeuds du sous arbre de racine v sont échangés entre les deux séquences.

Après avoir effectué ce crossover, une mutation est appliquée aux individus résultants, choisie parmi deux possibilités. La première est la mutation de type 2 vue dans la section 2.1. La seconde fonctionne de la manière suivante, en notant $p$ la longueur de la séquence et $\alpha$ un hyperparamètre à choisir :

— Avec probabilité $\frac{\alpha}{p}$ un noeud de l'arbre tiré uniformément est ajouté à la séquence .
— Avec probabilité $1 - \frac{\alpha}{p}$ un médicament de la séquence, tiré uniformément, est retiré.

**Critère d'arrêt** L'algorithme prend fin lorsque le critère est respecté. Dans notre cas, il s'agit d'un nombre d'itérations fixé par l'utilisateur.

# 3 Application et premiers résultats

## 3.1 Jeu de données simulé

Plusieurs jeux de données simulés ont été générés pour évaluer la performance de l'algorithme de Metropolis-Hastings. Pour ces jeux de données, plusieurs réponses sont définies préalablement, correspondant à divers cocktails de médicaments à risque. Chaque combinaison de médicaments à risque est d'intensité différente, c'est à dire, correspond à une probabilité plus ou moins faible de subir l'effet secondaire. Des informations ont étés recueillies auprès de pharmacologues dans le but de proposer un jeu de données simulé réaliste. Ce jeu de données comporte 200.000 patients. Il y a trois cocktails d'intérêt au sein de cette population. Chaque cocktail est pris par 1% de la population et provoque l'effet secondaire considéré avec une probabilité différente. Les trois probabilités sont $\frac{1}{50}$, $\frac{1}{100}$ et $\frac{1}{500}$. Un cocktail aléatoire donne l'effet secondaire d'intérêt avec probabilité $\frac{1}{2000}$.

## 3.2 Choix du score d'intérêt

Dans un premier temps. Une comparaison de la pertinence des différents indices de la section 1.3 est menée à l'aide de ces simulations.

Les différents indices ont été calculés sur 30 cocktails donnant lieu à un effet secondaire et 100.000 cocktails n'en donnant pas. Après classement par ordre croissant, cela permet de tracer des courbes Precision Recall présentées en Figure 4. Ces courbes sont identiques pour les trois méthodes RR, CRR et CSS, avec de très mauvais résultats. La méthode hypergéométrique donne de bien meilleurs résultats comme illustré également en Figure 5. En effet, les cocktails de plus haut score sont de vrais positifs, tandis qu'il s'agit de faux positifs pour les trois autres indices. Sur la Figure 5, les points de même abscisse sont légèrement décalés les uns par rapport aux autres sur l'axe des ordonnées de sorte à mieux visualiser les zones contenant beaucoup de cocktails.

L'analyse de disproportionnalité à travers les cocktails de médicaments est une tâche non triviale pour plusieurs raisons, l'une d'entre elles étant le bruit. Par bruit, on entend ici les ensembles de médicaments étant pris par peu de personnes dans le jeu de données. Certaines d'entre elles subissent l'effet secondaire étudié. Cela entraîne des valeurs de risques remarquablement élevées (voir CRR Figure 5), tandis qu'une conclusion sur la dangerosité du cocktail concerné semble impossible en raison de la taille de l'échantillon. De plus, le risque relatif attribue la même valeur à un cocktail consommé par 3 personnes, dont une

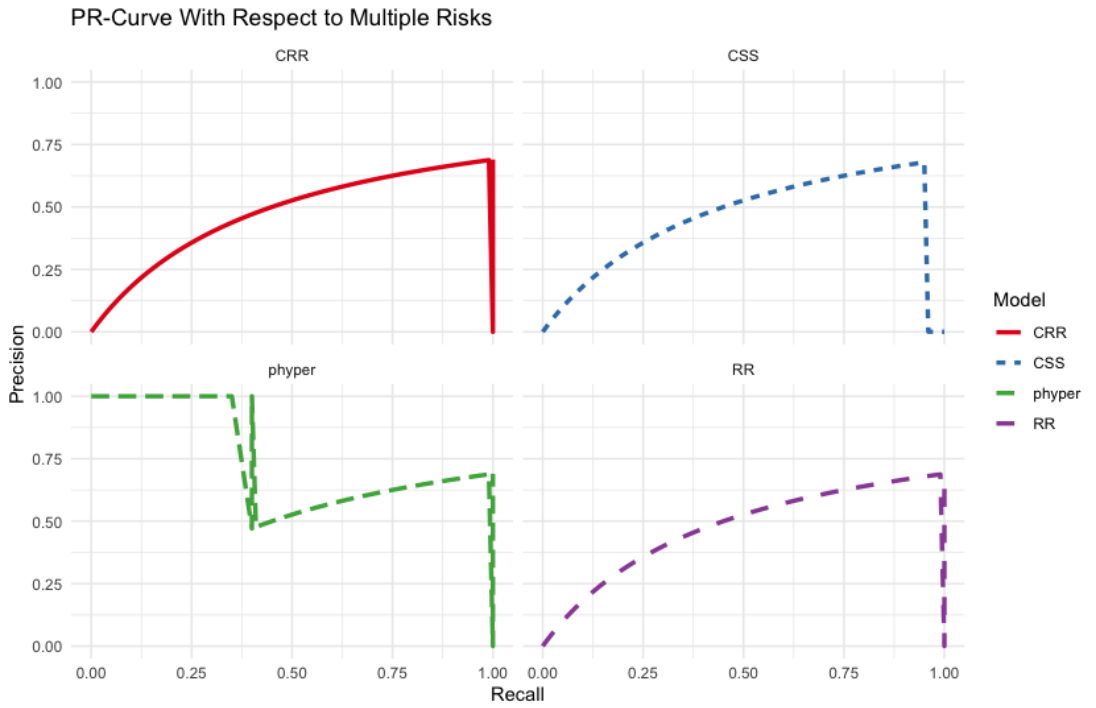



FIGURE 4 – Comparaison des courbes Precision-Recall associées aux différents risques pour un jeu de données simulé

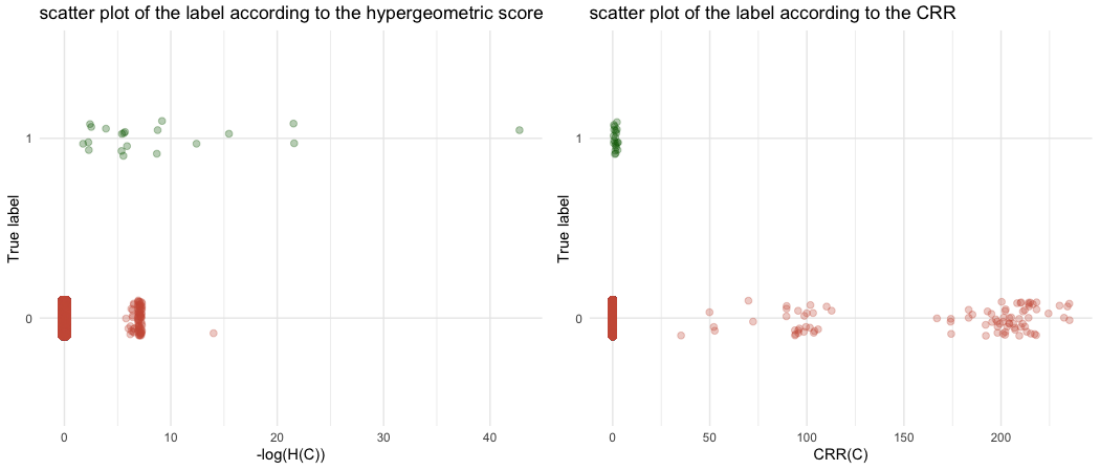



FIGURE 5 – Scatter plot montrant les scores des cocktails en fonction de leur vrai label, à gauche pour le score hypergéométrique, à droite pour le CRR

subit l'effet secondaire, qu'à un cocktail pris par 30 personnes, avec 10 d'entre elles subissant l'effet secondaire. Or, ce dernier est vraisemblablement plus à risque que le cocktail peu pris.

Le score hypergéométrique permet de prendre cet aspect en compte en attribuant un risque plus élevé aux combinaisons étant prises par un plus grand nombre de personne pour une même proportion d'effets secondaires observés.

## 3.3 Approximation de la distribution du risque à travers les cocktails de médicaments

L'algorithme MCMC permet d'obtenir une estimation de la distribution du risque à travers la population des cocktails d'une taille fixée. Dans notre exemple, les simulations sont effectuées sur des cocktails de taille 2 en utilisant le score hypergéométrique. Ainsi, il est possible de calculer la distribution réelle du risque en explorant de manière exhaustive toutes les combinaisons médicamenteuses de taille 2. Au delà, la combinatoire des cocktails devient trop grande pour calculer la distribution exhaustivement en un temps raisonnable.

La Figure 6 présente l'histogramme des deux distributions en ne gardant que les cocktails de risque non nul, ainsi que le QQ-plot associé. Les cocktails de risque nul sont éliminés car ils sont majoritaires au point d'écraser le reste de l'histogramme. L'approximation ainsi que la distribution réelle se trouvent respectivement en haut à gauche et en bas à gauche de la Figure 6. On constate que l'approximation en dimension deux est satisfaisante. On remarque de plus que la majorité de la distribution se situe aux alentours de 0.

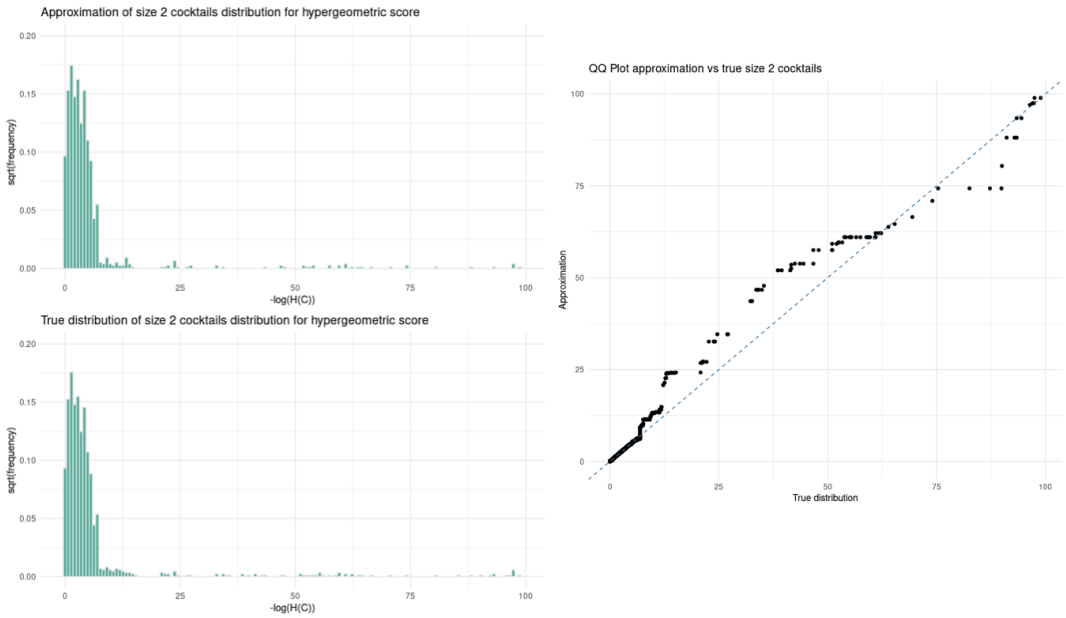



FIGURE 6 – Comparaison de la distribution réelle à la distribution estimée, conditionnellement à un risque non nul, à l'aide d'un diagramme quantile-quantile

Le QQ-plot n'est pas parfait mais il est à noter que 99% des scores correspondent à des points dans le segment initial en bas à gauche de la figure.

## 3.4 Algorithme génétique et données réelles

L'implémentation de l'algorithme génétique ainsi qu'une application des méthodes développées au jeu de données FAERS (Food & Administration (2024)), qui est un jeu de données publique de la Food & Drug Administration, sont en cours.

# Bibliographie


Bate, A., & Evans, S. (2009). Quantitative signal detection using spontaneous adr reporting. *Pharmacoepidemiology and drug safety*, *18*(6), 427–436.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. M. (1998). A bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, *54*, 315–321.

Evans, S. J., Waller, P. C., & Davis, S. (2001). Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety*, *10*(6), 483–486.

Food, & Administration, D. (2024). *Fda adverse event reporting system (faers) quarterly data extract files.* Consulté sur `https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html` ([En ligne ; Page disponible le 31/01/2024])

Goldberg, D. E. (2013). *Genetic algorithms*. pearson education India.

Maignen, F., Hauben, M., Hung, E., Holle, L. V., & Dogne, J.-M. (2014). A conceptual approach to the masking effect of measures of disproportionality. *Pharmacoepidemiology and drug safety*, *23*(2), 208–217.

Noguchi, Y., Aoyama, K., Kubo, S., Tachi, T., & Teramachi, H. (2020). Improved detection criteria for detecting drug-drug interaction signals using the proportional reporting ratio. *Pharmaceuticals*, *14*(1), 4.

Norén, G. N., Sundberg, R., Bate, A., & Edwards, I. R. (2008). A statistical methodology for drug–drug interaction surveillance. *Statistics in medicine*, *27*(16), 3057–3070.

van Puijenbroek, E. P., Bate, A., Leufkens, H. G., Lindquist, M., Orre, R., & Egberts, A. C. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, *11*(1), 3–10.

Walkup, J. T., Albano, A. M., Piacentini, J., Birmaher, B., Compton, S. N., Sherrill, J. T., ... others (2008). Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *New England Journal of Medicine*, *359*(26), 2753–2766.

Wikipédia. (2024). *Pharmacovigilance — wikipédia, l'encyclopédie libre.* Consulté sur `http://fr.wikipedia.org/w/index.php?title=Pharmacovigilanceoldid=204757946` ([En ligne ; Page disponible le 31/01/2024])

# Synthèse de données par la méthode Avatar : anonymisation et fidélité en pharmacologie de la transplantation

C. Benoist[1] & P. Marquet[1,2] & F. Stanke-Labesque[3] & J.-B. Woillard[1,2]

[1] *Service de pharmacologie, Toxicologie et pharmacovigilance, CHU de Limoges, France*

[2] *Pharmacologie & Transplantation, INSERM 1248, Université de Limoges, Limoges, France*

[3] *Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, HP2, Grenoble 38000, France*

Auteur correspondant : Clément Benoist, Clement.BENOIST@chu-limoges.fr

**Résumé.** La confidentialité est centrale dans l'utilisation des données médicales. Néanmoins, à l'heure des entrepôts de données de santé, cette contrainte freine l'exploitation des données médicales. La question de l'anonymisation est une question difficile. Le remplacement des noms par des pseudonymes ne suffit pas : il est nécessaire qu'on ne puisse pas réidentifier les patients de la base de données. Les données médicales étant coûteuses à produire, l'idée est de produire des données synthétiques à la fois anonymisées et fidèles aux données originales. L'algorithme Avatar vise à la production de données synthétiques anonymisées ; il est reconnu conforme en termes d'anonymisation par la Commission National de l'Informatique et des Libertés, sous conditions. Nous le comparerons à l'algorithme CT-GAN (conditionnal tabular generative adversial network). Nous avons décidé de considérer des métriques de fidélité (inverse de la divergence de Kullback-Leibner, $p$-valeur du test de Kolmogorov-Smirnov) et de confidentialité (valeur de k-anonymisation). Deux jeux de données seront utilisés dans le domaine de la transplantation rénale : un jeu de données relatif à l'effet de l'inflammation sur l'exposition au tacrolimus (médicament anti-rejet, servant à prévenir la réaction immunitaire contre le greffon) et un jeu de données de pharmacocinétique de population. Nous montrerons qu'Avatar semble avoir une bonne capacité d'anonymisation sur les jeux de données de taille moyenne (inflammation du tacrolimus), sans perdre en fidélité.

**Mots-clés.** Anonymisation, confidentialité, Avatar, données génératives, médecine

**Abstract.** Confidentiality is central to the use of medical data. However, in the age of health data warehouses, this constraint is holding back the exploitation of medical data. Anonymization is a difficult issue. Replacing names with pseudonyms is not enough : patients must not be re-identified in the database. As medical data is expensive to produce, the idea is to produce synthetic data that is both anonymized and faithful to the original data. The Avatar algorithm is designed to produce anonymized synthetic data ; it is recognized as compliant in terms of anonymization by the Commission National de l'Informatique et des Libertés, subject to certain conditions. We will compare it with the CT-GAN (conditional tabular generative adversial network) algorithm. We have decided to consider fidelity metrics (inverse of the Kullback-Leibner divergence, $p$-value of the Kolmogorov-Smirnov test) and confidentiality (k-anonymization value). Two datasets will be used in the field of renal

transplantation : a dataset relating to the effect of inflammation on exposure to tacrolimus (anti-rejection drug, used to prevent the immune reaction against the graft) and a population pharmacokinetics dataset. We'll show that Avatar appears to have good anonymization capability on medium-sized datasets (tacrolimus inflammation), without losing fidelity.

**Keywords.** Anonymisation, privacy, Avatar, generative data, medicine.

# 1 Introduction

La confidentialité des données médicales est particulièrement importante dans de nombreux domaines, par exemple en médecine : par exemple, la connaissance d'un diagnostic médical pourrait mener à une augmentation des primes d'assurance ou à une discrimination à l'embauche. Une première idée pour garantir la confidentialité est de supprimer les identifiants et de les remplacer par des pseudonymes. Néanmoins certaines variables sensibles , seules ou combinées, éventuellement croisées avec d'autres jeux de données, permettent de ré-identifier le patient. Par exemple, Culnane *et al.* [2017] montrent que, sur une base de données de santé australienne en open data (MBS/MPS), 10% de la base de données étaient potentiellement ré-identifiable. Certaines ré-identifications se fondaient sur les données rares (prescriptions ou maladies rares, grossesses multiples). Les données ne sont anonymisées que si on ne peut pas réidentifier les patients (sinon, elles sont pseudonymisées). D'un point de vue légal, le **r**èglement **g**énéral sur la **p**rotection des **d**onnées (RGPD) s'applique à partir du moment où les données sont pseudonymisées.

Ces considérations légitimes de vie privée ralentissent l'échange de données sensibles entre les acteurs utilisant les données de santé. Ainsi une idée sera de créer des données synthétiques répondant à deux impératifs concurrents : la fidélité qui exige que les distributions des données originales et des données synthétiques soient identiques et l'anonymisation qui requiert qu'on ne puisse déduire des informations supplémentaires sur les patients que ces patients aient servi à générer les données synthétiques ou non (en s'inspirant de la confidentialité différentielle). Dans les applications médicales, on peut considérer la fidélité quand les jeux de données synthétiques aboutissent aux mêmes résultats dans les analyses statistiques ; on peut aussi considérer des métriques qui indiquent le niveau de fidélité. De même, en médecine, on peut considérer qu'un jeu de données est anonyme si on ne peut pas ré-identifier les patients dans le sens courant ; de même, on peut calculer des métriques pour évaluer l'anonymisation d'un jeu de données.

Il existe des méthodes de synthèse de données similaires à un jeu de données telles que les GAN (generative adversial network), la méthode Synthpop, la méthode SMOTE (**s**ynthetic **m**inority **o**versampling **te**chnique) ou Stable diffusion. Néanmoins, ces méthodes ne sont pas orientées vers la confidentialité. La méthode Avatar [Guillaudeux et al., 2023] vise à l'anonymisation des données, elle a obtenu une certification de la CNIL (**c**ommission **n**ationale de l'**i**nformatique et des **l**ibertés), autorité française indépendante chargée de la vie privée et de la protection des données, en très grande majorité dans le domaine de l'informatique. Cette certification concerne la capacité de l'algorithme à anonymiser les données.

Les algorithmes d'anonymisation, qui génèrent des données, peuvent être également uti-

lisés pour augmenter les données.

# 2 Matériels et méthodes

## 2.1 Données

Deux jeux de données sont considérés :

- un jeu de données de pharmacocinétique de population [1] pour du tacrolimus en transplantation cardiaque. Le jeu de données comporte 29 observations et 18 variables. Il y a des variables continues et des variables catégorielles. Il y a des variables démographiques (âge, sexe), des données de concentrations du tacrolimus dans le sang.
- un jeu de données sur l'effet d'inflammation sur l'exposition au tacrolimus (NCT00812786) en transplantation hépatique ; l'exposition au tacrolimus correspond à sa courbe de concentration en fonction du délai avant la prise. Le jeu de données comporte 1573 observations. Il y a 11 variables, parmi lesquelles des marqueurs d'exposition du tacrolimus et des variables démographiques (age,sexe...), biologiques. Les données combinent des données catégorielles et continues.

L'objectif est de comparer les méthodes d'anonymisation sur des jeux de données de petit taille (ce qui est fréquent en médecine) et de grand taille (dans le domaine médical : la grande taille est relative).

## 2.2 Méthodes de génération de données synthétiques

### 2.2.1 Méthode à tester : Avatar

Notre méthode reprend l'algorithme d'Avatar. En entrée, nous avons un jeu de données pseudonymisées. Les données sont projetées dans un espace latent multidimensionnel par analyse en composantes principales. En considérant uniquement les $n_d$ premières composantes principales, on cherche les $k$ plus proches voisins, pour une donnée dans l'espace latent. Pour chaque individu, un avatar est synthétisé de manière stochastique à partir de ses $k$ plus proches voisins. Un nombre quelconque d'avatars peut être généré.

Considérons les $k$ plus proches voisins d'un individu $O$ dans l'espace latent, chacun sera associé à un coefficient. Soit $D_i$ l'inverse de la distance entre $O$ et le $i$-ème plus proche voisin. Soit $R_i$ une réalisation de la loi exponentielle d'intensité $\lambda$. Pour obtenir le vecteur $(C_i)_{i=1}^k$, on permute le $k$-uplet $(1/2, (1/2)^2, \ldots, (1/2)^k)$.

[1] La pharmacocinétique mesure l'exposition d'un médicament. L'exposition peut être représentée par la concentration du médicament en fonction du temps. Pour cela, on utilise des modèles paramétriques appelé modèles pharmacocinétiques. La variabilité sur les paramètres pharmacocinétiques en fonction des individus et de leur caractéristiques associées (sexe, âge...), souvent à l'aide de modèles non linéaires à effets mixtes, est le champ d'étude de la pharmacocinétique.

Le poids $P_i$ sera calculé de manière naturelle : $P_i = D_i \times R_i \times C_i$.

Le poids sera normalisé :

$$W_i = \frac{P_i}{\sum_{j=1}^{k} P_j}$$

Ainsi, $W_i$ sera le poids associé au $i$-ème voisin. Une combinaison linéaire permet d'obtenir les avatars.

L'opération peut être répétée.

### 2.2.2 Méthode de référence : CTGAN (conditionnal tabular generative adversial network)

L'algorithme CTGAN décrit dans [Xu *et al.*, 2019] est une méthode qui, à partir d'un apprentissage, génère des données synthétiques tabulaires similaires aux données d'apprentissage. Je rappelle le détail de cet algorithme.

On considère une table où $N_c$ colonnes $C_1, \dots, C_{N_c}$ correspondent aux réalisations de variables aléatoires continues et $N_d$ colonnes $D_1, \dots, D_{N_d}$ correspondent aux réalisations de variables aléatoires discrètes.

La première étape est de prétraiter les colonnes associées à des variables continues grâce à une normalisation à mode spécifique. Les colonnes associées à des variables aléatoires continues notées $C_i = (c_{i,j})_{i,j}$ seront modifiées en utilisant le modèle à mélange de gaussiennes variationnel : le nombre de modes $m_i$ est estimé et le mélange de gaussiennes est ajusté pour en déterminer les paramètres. Notons ces modes $\eta_k$, $k = 1, \dots, m_i$. Ainsi la densité de $C_i$ peut s'écrire

$$f_{C_i} = \sum_{k=1}^{m_i} \mu_{i,j,k} \mathcal{N}(c_{i,j}; \eta_{i,j,k}, \varphi_{i,j,k})$$

où $\mathcal{N}(\cdot; \tilde{\eta}, \tilde{\varphi})$ est la densité de la loi normale de moyenne $\tilde{\eta}$ et d'écart-type $\tilde{\varphi}$.

Pour chaque valeur de probabilité, on calcule les densités de probabilité des $c_{i,j}$ associées à chacun de leur mode. On a des densités $\rho_{i,j,1}, \dots, \rho_{i,j,m_i}$. Ces densités valent $\rho_{i,j,k} = \mu_k \mathcal{N}(c_{i,j}; \eta_k, \varphi_k)$.

On cherche $\ell^*_{i,j}$ qui maximise en $\ell$ la quantité $\rho_{i,j,\ell}$ ; notons ce maximum $\rho^*_{i,j}$. Nous considérons $c_{i,j,\ell^*_{i,j}}$. Nous posons $\beta_{i,j}$ le vecteur de taille $m_i$ qui vaut 0 partout sauf dans la position d'ordre $\ell^*_{i,j}$ où la valeur vaut 1. On considère aussi le scalaire $\alpha_{i,j} = \frac{c_{i,j,\ell^*_{i,j}} - \eta_{i,j,\ell^*_{i,j}}}{4\varphi_{\ell^*_{i,j}}}$.

La ligne sera représentée par :

$$r_j = \alpha_{1,j} \oplus \beta_{1,j} \oplus \cdots \oplus \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus d_{1,j} \oplus \cdots \oplus d_{N_d,j}$$

où $\oplus$ est l'opérateur de concaténation, $N_c$ (respectivement $N_d$) est le nombre de variables aléatoires continues (respectivement discrètes) et les $d_{i,j}$ sont encodés selon l'encodage One-Hot des colonnes correspondant aux variables aléatoires discrètes.

Le GAN utilise deux réseaux de neurones : un générateur, qui génère des données synthétiques et un discriminateur, qui propose un score qui évalue la proximité statistique entre les vraies données et les données synthétiques, issues du générateur (conditionnel).

Le CTGAN utilise un générateur conditionnel. Soit $i^*$ un indice de numérotation des colonnes discrètes. Soit $R$ la variable aléatoire sous-jacente à la génération d'une ligne.

Nous considère le vecteur conditionnel *cond* relatif à la condition $(D_{i*} = k^*)$. Soit $i \in [\![1, N_d]\!]$. Soit $m_i = (m_i^{(k)})_{k=1}^{N_d}$ où $m_i^{(k)}$ vaut 0 sauf si $i = i^*$ et $k = k^*$ où il vaut 1. Ce vecteur conditionnel conditionnel, concaténé à des réalisations de la loi normale centrée réduite regroupée en vecteur sert d'entrée au générateur.

Il est fourni par le générateur conditionnel $\{\hat{d}_1, \ldots, \hat{d}_{N_d}\}$. La fonction de perte est pénalisée en ajoutant la cross-entropy entre $m_{i^*}$ et $\hat{d}_{i^*}$

L'entraînement, appelé entraînement par échantillon, s'effectue de la manière suivante :

- On pose les $N_d$ vecteurs remplis de zéros $m_i = (m_i^{(k)})_{i=1}^{|D_i}$.
- On sélectionne aléatoirement un $i^*$ selon une loi uniforme discrète sur $[\![1, N_d]\!]$, qui est associés à $D_{i^*}$.
- On élabore une fonction de pondération fondée sur la colonne $D_{i^*}$
- On tire $k^*$ de la fonction de pondération précitée
- On constitue le vecteur conditionnel correspondant.

Dans l'implémentation du CTGAN et le calcul des métriques, présentées ci-après, nous avons le package *synthcity* [Zhaozhi *et al.*, 2023].

## 2.3 Métriques

Il y a plusieurs métriques : certaines évaluent la fidélité des données, certaines évaluent la confidentialité.

La première métrique est la p-valeur dans le test de Kolmogorov-Smirnov. Elle est comprise entre 0 et 1. Plus cette métrique est proche de 1, plus les distributions considérées (en particulier, entre données réelles et synthétiques) sont proches.

Une autre mesure de fidélité est la divergence estimée de Kullback-Leibler. Plus cette mesure est proche de 0, plus les distributions (par exemple, entre données réelles et synthétiques) sont proches. Dans notre cas, nous présenterons l'inverse de la distribution de Kullback-Leibler ; dans ce cas, plus la valeur de cet inverse est proche de 0, plus les distributions (par exemple, entre données réelles et synthétiques) sont différentes.

La mesure de la $k$-anonymisation est l'effectif minimum dans les catégories, qui sont reconstituées par l'algorithme des $k$-means.

Ces mesures sont obtenues par le package *synthcity* [Qian *et al.*, 2023].

| | Avatar knn=5 | Avatar knn=10 | Avatar knn=50 | CT-GAN |
|---|---|---|---|---|
| KL-inverse | 0.756/0.972 | 0.856/0.930 | 0.775/0.972 | 0.774/0.759 |
| KS test | 0.927/0.930 | 0.903/0.909 | 0.872/0.930 | 0.846/0.853 |
| $k$-anonymisation | 24 (origine : 12)/ 91 (origin : 12) | 29/119 | 25/91 | 17/52 |

TABLE 1 : Métriques comparant le jeu de données initiales et le jeu de données synthétiques, pour le jeu de donnée sur l'inflammation par le tacrolimus. La première valeur correspond au jeu de données synthétiques de même taille que le jeu de données intiales, la seconde valeur correspond à un jeu de données synthétique qui fait 4 fois la taille des données initiales.

| | Avatar knn=5 | Avatar knn=10 | CT-GAN |
|---|---|---|---|
| KL-inverse | 0.358/0.616 | 0.226/0.350 | 0.364/0.532 |
| KS test | 0.822/0.847 | 0.78/0.802 | 0.713/0.705 |
| $k$-anonymisation | 11 (orig. : 14)/ 3 (orig. : 14) | 10/5 | 11/3 |

TABLE 2 : Métriques comparant le jeu de données initiales et le jeu de données synthétiques, pour le jeu de donnée de pharmacocinétique de population. La première valeur correspond au jeu de données synthétiques de même taille que le jeu de données intiales, la seconde valeur correspond à un jeu de données synthétique qui fait 4 fois la taille des données initiales.

### 2.4 Plan de travail

Les données synthétiques par les algorithmes de génération de données suivants :

- l'algorithme Avatar avec un nombre de plus proches voisins égal à 5
- l'algorithme Avatar avec un nombre de plus proches voisins égal à 10
- l'algorithme Avatar avec un nombre de plus proches voisins égal à 50, pour le jeu de données d'inflammation du tacrolimus uniquement
- CT-GAN

Pour chacun de ces algorithmes de génération de données, deux jeux de données seront synthétisés : un de même taille que le jeu de données original, l'autre qui fait 4 fois la taille du jeu de données original.

## 3 Résultats

Les tables 1 et 2 présentent les métriques de fidélité et de confidentialité relatives respectivement au jeu de données d'inflammation par le tacrolimus et au jeu de données de pharmacocinétique de population.

# 4 Discussion

L'utilisation de deux jeux de données permet de tester l'efficacité de la méthode de synthèse Avatar sur un petit jeu de données (pharmacocinétique) et sur un jeu de données plus conséquent (inflammation du tacrolimus).

Pour le jeu de données d'inflammation (tacrolimus), on constate une amélioration des métriques de confidentialité : la valeur de k-anonymisation est plus grande que la valeur d'origine, quel que soit l'algorithme de synthèse utilisé. Cette amélioration se renforce lorsque l'on génère plus de données. On constate que Avatar est plus performant en termes de confidentialité que le CTGAN. On constate que, sur la valeur de k-anonymisation, le nombre optimal de plus proches voisins utilisé dans Avatar est 10, ce qui montre que le nombre de plus proches voisins est optimisable.

Pour le jeu de données de pharmacocinétique, on constate une détérioration de la valeur de k-anonymisation lorsque l'on utilise Avatar, détérioration qui se confirme lorsque l'on génère plus de données. Cette détéorioration s'explique vraisemblablement par le manque de données initiales ; la dégradation s'observe aussi avec CT-GAN avec approximativement la même intensité. Le nombre optimal de plus proches voisins dans Avatar sur la valeur de k-anonymisation n'est pas clair : d'un point de vue strictement numérique, il dépend de la taille de la l'échantillon généré, mais cette différence ne semble pas pertinente.

En ce qui concerne la fidélité des données synthétiques aux données réelles, les deux métriques utilisées sont : l'inverse de la divergence de Kullback-Leibner et la p-valeur de test de Kolmogorov-Smirnov.

Pour le jeu de données d'inflammation (tacrolimus), du point de vue de la $p$-valeur du test de Kolmogorov-Smirnov, pour une génération de données synthétiques de la taille du jeu de données initiales par Avatar, lorsque le nombre de plus proches voisins diminue, la $p$-valeur augmente, ce qui indique une fidélité des données générées aux données réelles qui s'améliore. Pour une génération de données de taille quatre fois, la taille des données initiale, la $p$-valeur du test de Kolmogorov-Smirnov reste globalement constante, ce qui semble indiquer que la fidélité ne dépend pas du nombre de voisins retenu.

Si on compare, en termes de $p$-valeur du test de Kolmogorov-Smirnov, la fidélité relative aux méthodes Avatar est meilleure que celle relative aux méthodes CT-GAN, ce qui est d'autant plus vrai que la quantité de données synthétiques est grande.

Toujours pour le jeu de données d'inflammation (tacrolimus), en ce qui concerne l'inverse de la divergence de Kullback-Leibner, la méthode Avatar est au moins aussi performante que la méthode CT-GAN (sauf d'un point de vue numérique, dans un cas) ; la fidélité s'améliore, lorsque l'on augmente la taille des données générées, pour les méthodes Avatar ; cette amélioration est plus modeste pour CT-GAN.

Pour le jeu de données de pharmacocinétique, l'augmentation de la taille des données synthétiques améliore la fidélité du point de vue de l'inverse de la divergence de Kullback-Leibner et de la $p$-valeur du test de Kolmogorov-Smirnov. Dans Avatar, l'utilisation de 5 plus proches voisins aboutit à une meilleure fidélité que l'utilisation de 10 plus proches voisins ; et selon ces deux métriques, Avatar est globalement et approximativement aussi performant

que CT-GAN.

Notons qu'il y a probablement, dans le jeu de données de pharmacocinétique, une variabilité des métriques (en particulier de la p-valeur du test de Kolmogorov-Smirnov) du fait de la faible taille des données.

D'une manière générale, la capacité d'anonymisation de Avatar semble nécessiter un volume de données relativement conséquent, ce qui devra être confirmée par d'autres études. Cependant, la génération de données médicales coûte cher et les jeux de données sont par conséquent réduits. Trouver un algorithme qui génère des données synthétiques, à la fois fidèles et anonymisées, reste un défi.

Comme perspective, on peut vérifier que les données synthétiques engendrent des résultats statistiques similaires à ceux obtenus par des données réelles.

# Remerciements

Ce travail fait partie du projet DIGPHAT, qui est soutenu par le gouvernement français, dirigé par l'agence nationale de la recherche (ANR) dans le cadre du programme France 2030 (référence : ANR-22-PESN-0017).

# References

Culnane, C., Rubinstein, B. I., & Teague, V. (2017). Health data in an open world. *arXiv preprint arXiv :1712.05627*.

Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C. A., Goronflot, T., ... & Gourraud, P. A. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, 6(1), 37.

Qian, Z., Cebere, B. C., & van der Schaar, M. (2023). Synthcity : facilitating innovative use cases of synthetic data in different data modalities. *arXiv preprint arXiv :2301.07573*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32.

# Réseaux de neurones 1

# Estimation de durée de vie de filtres moteurs en prenant en compte les données censurées

Jean-Pierre Noot [1,2] & Etienne Birmelé [1] & François Rey [2]

[1] *Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS*
*7 rue René-Descartes, 67000 Strasbourg, France,*
*jnoot@unistra.fr*
[2] *Liebherr-Components Colmar SAS, 68025 Colmar, France,*
*jean-pierre.noot@liebherr.com*

**Résumé.** La maintenance prédictive consiste à anticiper les défaillances des composants industriels pour effectuer le remplacement de ces composants au meilleur moment. Elle permet de prévenir les arrêts, comme dans la maintenance réactive, et de réduire les coûts par rapport à la maintenance préventive. L'estimation de la durée de vie restante (Remaining Useful Life RUL) des composants industriels est devenue un défi majeur pour la maintenance prédictive. Dans de nombreuses applications, en particulier celles pour lesquelles la maintenance préventive est la règle générale, le problème de prédiction de durée de vie restante est rendu difficile par la rareté des instances défaillantes. En effet, l'interruption de l'acquisition de données avant la survenue de l'événement d'intérêt conduit à des données censurées à droite. Il est courant dans le milieu industriel d'avoir un taux élevé de données censurées.

Ce document propose une approche d'apprentissage profond basée sur des séries temporelles issues de plusieurs capteurs, qui permet de prendre en compte les données censurées lors de l'entraînement des réseaux de neurones. Pour ce faire, le problème est discrétisé, comme dans Vishnu, Malhotra, Vig, & Shroff (2019), l'objectif étant de prédire un vecteur binaire dont les coordonnées correspondent à l'état du composant à des moments prédéfinis dans le futur (1, défaillant ; 0, sain). Une architecture de réseau de neurones est proposée, tirée de Dual Aspect Self-Attention based on Transformer (DAST) Zhang, Song, & Li (2022). Cette architecture a été initialement développée pour faire de la régression sur des données non censurées. L'architecture de DAST prend en compte les dépendances temporelles à court terme ainsi que les interactions entre capteurs. Deux méthodes sont présentées dans ce document, l'estimation directe du RUL avec DAST et l'estimation indirecte du RUL prenant en compte la censure avec DAST-OR.

DAST et DAST-OR ont été validés sur le jeu de données public du C-MAPSS. Puis ils ont été mis en place sur une application Liebherr, l'objectif étant de développer des modèles capables d'estimer la durée de vie restante de filtres moteurs, en particulier des filtres à huile. Une proportion importante des données disponibles de filtres à huile Liebherr sont censurées, d'où le développement de la méthode DAST-OR permettant de prendre la censure en compte.

**Mots-clés.** Maintenance prédictive, estimation de durée de vie, apprentissage statistique, apprentissage profond.

**Abstract.** Predictive maintenance deals with the timely replacement of industrial components relatively to their failure. It allows to prevent shutdowns as in reactive maintenance and

reduces the costs compared to preventive maintenance. As a consequence, Remaining Useful Life (RUL) prediction of industrial components has become a key challenge for condition based monitoring. In many applications, in particular those for which preventive maintenance is the general rule, the prediction problem is made harder by the rarity of failing instances. Indeed, the interruption of data acquisition before the occurrence of the event of interest leads to right censored data. It is common in the industrial environment to have a high rate of censored data.

The present document proposes a deep-learning approach based on multi-sensor time series which allows to consider censored data during the training of the neural networks. To do so, the problem is discretised, following Vishnu, Malhotra, Vig, & Shroff (2019), into the prediction of a vector whose coordinates correspond to the status of the component at pre-defined moments in the future (1, failed ; 0, healthy). A neural network architecture is proposed, based on the Dual Aspect Self-Attention based on Transformer (DAST) from Zhang, Song, & Li (2022). The DAST architecture take into account short-term temporal dependencies as well as interactions between sensors. Two methods are presented in this document, the direct RUL estimation with DAST and the indirect RUL estimation which takes into account censored data with DAST-OR.

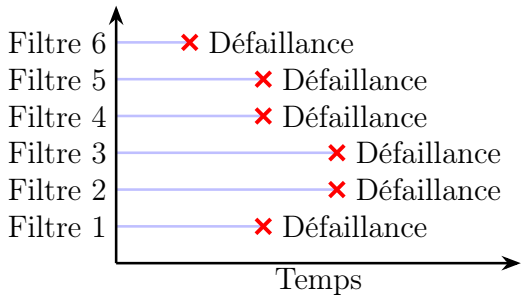
DAST and DAST-OR have been implemented on a Liebherr application with the objective of developing models capable of estimating the remaining useful life (RUL) of motor filters, particularly oil filters. An important proportion of available data on Liebherr oil filters are censored, hence the development of the DAST-OR method to take into account censored data for the training.

**Keywords.** Predictive maintenance, remaining useful life estimation, statistical learning, deep learning.

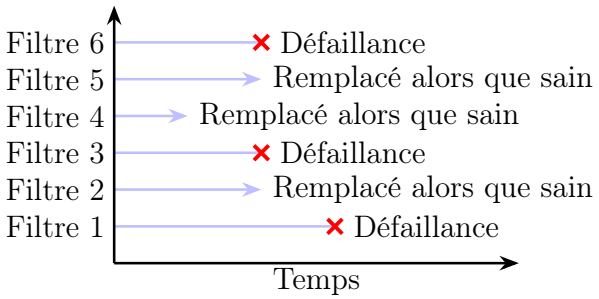
# 1 Introduction

Liebherr est un fabricant d'équipements composé de plus de 130 entreprises, organisées en plusieurs divisions telles que le terrassement, l'exploitation minière, les appareils électroménagers, les composants moteurs, etc. L'usine de Liebherr Components Colmar (COC) produit des moteurs diesel haute puissance ($> 1$ MW). C'est dans ce cadre que des recherches sont menées pour créer des algorithmes de maintenance prédictive sur les composants moteurs. Les avancées technologiques et électroniques des capteurs de nos jours permettent la collecte d'une quantité considérable de données sur les équipements mécaniques et industriels comme les moteurs produit par la COC, en particulier des séries temporelles mesurant leur évolution au fil du temps. La définition du calendrier de maintenance, cruciale pour l'industrie, tend vers une maintenance prédictive, ou maintenance basée sur l'état (Condition-Based Monitoring CBM). Cette dernière est définie par opposition à la maintenance préventive historique, pour laquelle le calendrier de maintenance est prédéfini, chaque composant étant remplacé à des intervalles de temps fixes. La CBM évite le remplacement de composants sains, en établissant un programme dynamique qui évolue en fonction de la surveillance en temps réel du système,

réduisant ainsi les coûts. Une étape cruciale est donc l'estimation, compte tenu de l'utilisation réelle du système, de la durée de vie utile restante d'un composant, c'est-à-dire le temps avant sa défaillance.




(a) Données non censurées : toutes les durées de vie sont complètes ie la défaillance de tous les filtres est observée




(b) Données censurées à droite, les filtres 1, 3 et 6 sont censurés à droite, ils étaient encore sain lors du replacement

FIGURE 1 – Données non censurées et données censurées à droite

L'objectif est d'estimer la durée de vie restante (Remaining Useful Life, RUL) des filtres moteurs. Nous étudions une variable continue positive, la différence de temps entre le présent et le moment de l'événement d'intérêt, lorsque le filtre est obstrué.

À la COC, les filtres à huile sont remplacés avant d'être bouchés suivant le calendrier de maintenance pré-définis amenant à une fraction importante de données censurées à droite. Nous n'avons donc pas la durée de vie complète des filtres, mais une information partielle cf 1b. L'objectif est de créer un modèle pour estimer le RUL des filtres en considérant les données censurées.

# 2 État de l'art

## 2.1 Dual Aspect Self-Attention based on Transformer (DAST)

Le modèle DAST (Dual Aspect Self-Attention based on Transformer) a été introduit par Zhang, Song, & Li (2022) pour traiter des séries temporelles provenant de capteurs, le modèle repose sur l'architecture des transformateurs introduit dans Vaswani et al. (2017) avec quelques adaptations faites au modèle. L'entrée $X$ est une fenêtre glissante de taille $W$ des séries temporelles provenant des $p$-capteurs. Les entrées du modèle sont des matrices $X \in \mathcal{M}_{p,W}$.

En plus de la fenêtre glissante, la valeur moyenne et le coefficient de la régression linéaire en fonction du temps de chaque capteur sont ajoutées à $X$, enfin :

$$X(t) = \begin{pmatrix} x_{1,t} & \cdots & x_{1,t+W} & \mu_1(t) & \alpha_1(t) \\ \vdots & \ldots & \vdots & \vdots & \vdots \\ x_{p,t} & \cdots & x_{p,t+W} & \mu_p(t) & \alpha_p(t) \end{pmatrix} \tag{1}$$

où $p$ est le nombre de capteurs, $W$ est la taille de la fenêtre glissante, $x_{i,t}$ correspond à la valeur du capteur $i$ au moment $t$, $\mu_i$ et $\alpha_i$ sont la valeur moyenne et le coefficient de régression du capteur $i$ dans la fenêtre glissante. Finalement, $X \in \mathcal{M}_{p,W+2}$. Le traitement des données décrit dans Song et al. (2020) et utilisé dans Zhang, Song, & Li (2022) est utilisé.

La particularité de l'architecture DAST réside en la présence de deux encodeurs, **encodeur capteur** et **encodeur temporel**. Les encodeurs permettent d'apprendre simultanément les poids des différents capteurs et des pas de temps. Cette information est importante pour l'estimation de la durée de vie restante basée sur des séries temporelles provenant de capteurs. D'un côté la matrice enrichie $X(t)$ est donnée en entrée de l'encodeur temporel et de l'autre côté sa transposée $X(t)^T$ est donnée en entrée de l'encodeur capteur.

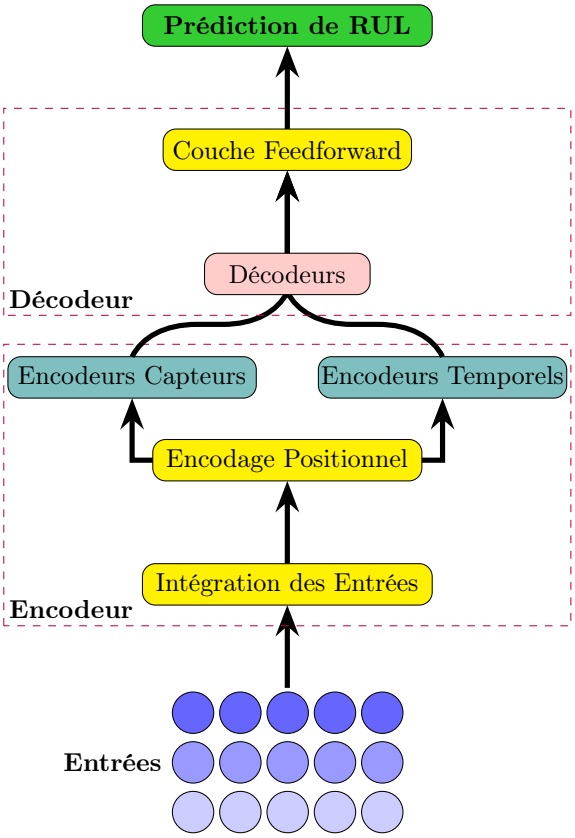



FIGURE 2 – Architecture originale de DAST Zhang, Song, & Li (2022)

## 2.2 Régression ordinale pour l'estimation de durée de vie restante avec des données censurées

Dans cette section, une méthode pour estimer le RUL des filtres avec des données censurées à droite et non censurées est détaillée. Elle repose sur une régression ordinale (OR) qui discrétise le problème. Au lieu d'estimer directement le RUL des filtres, celui-ci est transformé en un vecteur binaire de l'état du composant dans le futur. Pour cela à un temps fixé $t$ on défini une nouvelle cible $\hat{Y}_t = (y_{t,0}, \cdots, y_{t,n})$ où

$$y_i = \begin{cases} 1 & \text{si le filtre est bouché après } l * i \text{ heures,} \\ 0 & \text{si le filtre est sain après } l * i \text{ heures,} \\ \text{-} & \text{si l'état du filtre est inconnu après } l * i \text{ heures,} \end{cases}$$

pour $i \in \{1, \cdots, n\}$ où $n$ est la longueur de $\hat{Y}$ et $l$ une constante définie en fonction du problème étudié. Par exemple pour $n = 10$ et $l = 10$ heures :

— si le composant tombe en panne après $t + 75$ heures, $\hat{Y}_t = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1)$,

— si le composant est remplacé après $t + 75$ heures et est toujours en sain, $\hat{Y}_t = (0, 0, 0, 0, 0, 0, 0, \text{-}, \text{-}, \text{-})$. Les trois derniers éléments sont masqués.

L'erreur calculée est l'entropie croisée binaire (BCELoss), cette fonction d'erreur est largement utilisée pour les problèmes de classification binaire. Cette erreur sera ajustée pour

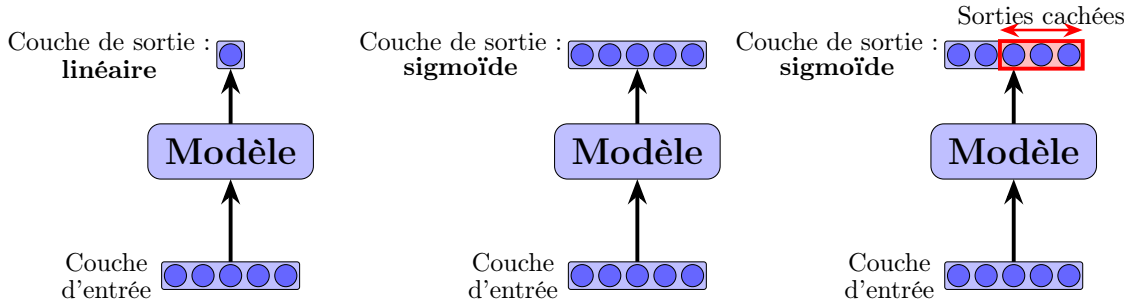
les données censurées à droite et ne sera calculée que sur $\hat{Y}_t$ partiel.

Si $\hat{Y}_t = (0, 0, 0, 0, 0, 0, 0, \text{-}, \text{-}, \text{-})$, la perte est calculée sur $\hat{Y}' = (0, 0, 0, 0, 0, 0, 0)$.

$$BCE(Y_t, \hat{Y}_t) = -\sum_{i=1}^{n'} \left( Y_{t,i} log(\hat{Y}_{t,i}) + (1 - Y_{t,ki}) log(1 - \hat{Y}_{t,i}) \right) \tag{2}$$

où $n' = n$ pour les instances défaillantes, et $n' = n - n_c$ pour les instances censurées, avec $n_c$ le nombre d'éléments masqués de $\hat{Y}$.

Dans cet article, une analyse comparative est réalisée entre le modèle DAST proposé dans Zhang et al. (2022) et le modèle DAST-OR, modèle DAST auquel on a ajouté une couche sigmoïde en sortie afin d'obtenir un vecteur de probabilité.




(a) Régression (b) Régression ordinale (c) Régression ordinale avec des données censurées

FIGURE 3 – Régression comparé à la régression ordinale

# 3 Résultats expérimentaux, estimation de la durée de vie restante des filtres à huile

## 3.1 Résultats sur les données C-MAPSS

### 3.1.1 Présentation des données C-MAPSS

La performance de la méthode proposée est évaluée sur l'ensemble de données C-MAPSS (Commercial Modular Aero Propulsion System Simulation), qui est utilisé comme référence pour les méthodes d'estimation de RUL. Cet ensemble de données simule des trajectoires de défaillance de moteurs de turbofan Saxena et al. (2008) avec des conditions de fonctionnement et modes de défaillance différents, conduisant à quatre sous-jeux de données, FD001, FD002, FD003 et FD004. Les caractéristiques des quatre jeux de données sont résumées dans le Tableau 1.

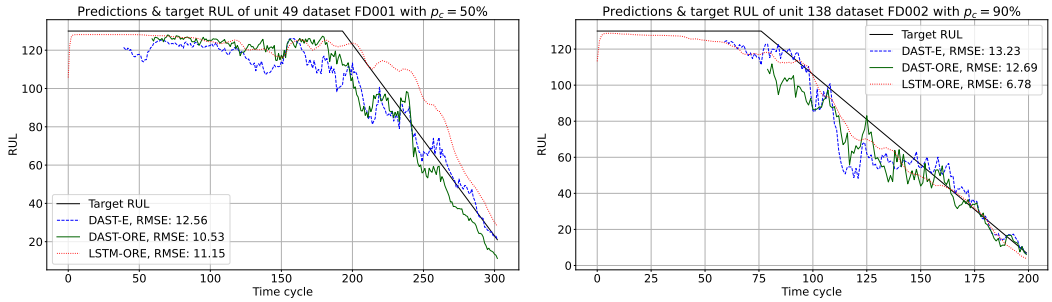
Chaque trajectoire contient les variables suivantes :

TABLE 1 – Présentation données C-MAPSS

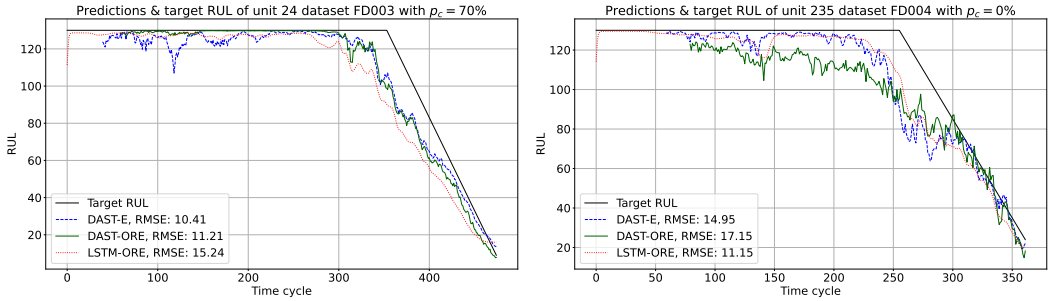
| Sous jeu de données | FD001 | FD002 | FD003 | FD004 |
|---|---|---|---|---|

1. un numéro d'unité correspondant à l'identifiant du composant,
2. une variable temporelle correspondant au nombre de cycles effectués,
3. les paramètres de simulation (condition de fonctionnement et modes de défaillance),
4. les données simulées provenant de 21 capteurs.

### 3.1.2 Résultats expérimentaux sur les données C-MAPSS




(a) Unité 49 (FD001) avec $p_c = 50\%$ (b) Unité (FD002) avec $p_c = 90\%$




(c) Unité 24 (FD003) avec $p_c = 70\%$ (d) Unité 235 (FD004) avec $p_c = 0\%$

FIGURE 4 – Exemple de résultats (RMSE) sur une unité de chaque sous-jeu du C-MAPSS

Trois modèles sont entraînés sur les jeux de données du C-MAPSS :
— DAST Zhang, Song, & Li (2022),
— LSTM-MLP-OR, amélioration que nous avons proposée de LSTM-ORCE de Vishnu, Malhotra, Vig, & Shroff (2019),
— DAST-OR.

Plusieurs entraînements (10) ont été réalisés pour chaque méthode, afin de réduire la part d'aléatoire dû à l'initialisation des poids des réseaux de neurones. Sur chacun des 10 entraînements réalisés, un modèle moyen est considéré, celui-ci correspond à la prédiction

moyenne faîtes par les six meilleurs modèles parmi les dix comme Vishnu, Malhotra, Vig, & Shroff (2019). De la censure est ajoutée artificiellement dans les données C-MAPSS pour évaluer les performances des modèles en fonction de l'évolution de celle-ci. Le taux de censure noté $p_c \in [0\%, 20\%, 50\%, 70\%, 90\%]$.

Lorsque il n'y a pas de censure dans les données, les résultats obtenus avec DAST-ORE sont comparables à ceux obtenus avec DAST et aux résultats que l'on retrouve dans la littérature cf Noot, Rey, & Birmelé (2024). DAST-ORE donne les meilleurs résultats sur FD001 et FD003, tandis que LSTM-MLP-ORE donne les meilleurs résultats sur FD002 et FD004. Lorsque de la censure est ajoutée manuellement dans les données, les résultats sont bien meilleurs que ceux obtenus dans la littérature, amélioration d'environ 20% sur FD001 et d'environ 100% sur FD004 cf Noot, Rey, & Birmelé (2024).

Le modèle DAST est performant sur les données du CMAPSS Zhang et al. (2022). Le modèle DAST-OR a également été validé sur ces données. De plus les données du C-MAPSS sont semblables aux données à notre disposition de part leur structure et la source de données (ce sont des série chronologiques issues de capteurs d'un système industriel). C'est pourquoi nous allons évaluer ces méthodes sur les données Liebherr.

TABLE 2 – Résultats (RMSE) sur les données FD001 et FD002 (les meilleurs résultats sont en gras)

| | **FD001** | | | | | |
|---|---|---|---|---|---|---|
| $p_c$ | LSTM-MLP-OR | LSTM-MLP-ORE | DAST | DAST-E | DAST-OR | DAST-ORE |
| 0% | 14.24 | 13.20 | 12.35 | 12.22 | 12.16 | **11.57** |
| 20% | 15.42 | 14.01 | 13.69 | 12.59 | 12.73 | **12.51** |
| 50% | 15.09 | 15.96 | 15.41 | 13.37 | 13.39 | **12.99** |
| 70% | 17.83 | 17.97 | 15.38 | 14.08 | 14.28 | **12.51** |
| 90% | 30.02 | 26.76 | 16.78 | 17.17 | 17.01 | **15.80** |
| | **FD002** | | | | | |
| $p_c$ | LSTM-MLP-OR | LSTM-MLP-ORE | DAST | DAST-E | DAST-OR | DAST-ORE |
| 0% | **12.00** | 12.77 | 16.48 | 15.44 | 15.62 | 15.55 |
| 20% | 15.43 | **13.01** | 14.09 | 13.80 | 16.37 | 18.51 |
| 50% | 13.71 | **13.15** | 15.08 | 14.18 | 15.39 | 16.58 |
| 70% | 14.24 | **13.24** | 16.10 | 14.74 | 16.71 | 17.73 |
| 90% | 16.44 | **13.61** | 15.85 | 15.08 | 25.23 | 17.00 |

## 3.2 Résultats sur les données Lieherr

### 3.2.1 Présentation des données Liebherr

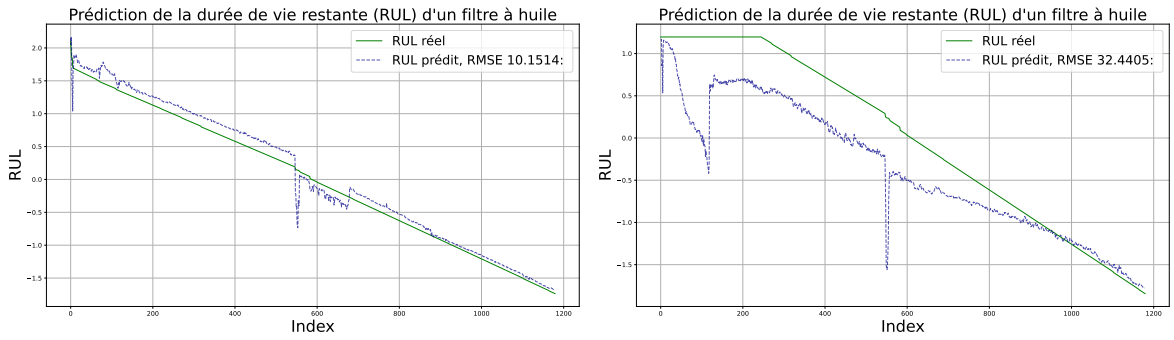
Les données utilisées sont issues des capteurs standards présents sur le moteur, tel que la vitesse moteur, l'injection, la pression, la température... Ces données sont échantillonnées à 1Hz. La détérioration des filtres étant un phénomène lent, les données sont moyennées, puis centrées réduites.

TABLE 3 – Présentation données Liebherr du banc d'essai

| | |
|---|---|
| Trajectoires disponibles | **11** |
| Trajectoires non censurées disponibles | **7** |
| Trajectoires censurées disponibles | **4** |

Une des difficultés principales de ce problème réside dans la faible quantité de données disponibles au banc d'essai. Il y a peu de filtres pour lesquelles les données complètes sont à disposition et pour lesquels l'état des filtres est connu (7) cf tableau 3. D'où l'intégration des données censurées dans l'entraînement des modèles afin d'augmenter la quantité de données à notre disposition. Néanmoins la quantité de données Liebherr disponible est faible (de 7 à 11 trajectoires) par rapport aux données du C-MAPSS (minimum 100 trajectoires). Cela ne permet pas de séparer les données Liebherr en trois jeux indépendants (entraînement, validation, test). DAST requérant des durées de vie complètes, seules les données des 7 filtres non censurés sont utilisées. Au contraire, pour DAST-OR toutes les données disponibles sont utilisées, soit 11 filtres.

### 3.2.2 Résultats expérimentaux sur les données Liebherr




(a) Résultats de DAST sur le filtre à huile 1 (b) Résultats DAST-OR sur le filtre à huile 1

FIGURE 5 – Résultats de DAST et DAST-OR pour l'estimation directe du RUL du filtre 1

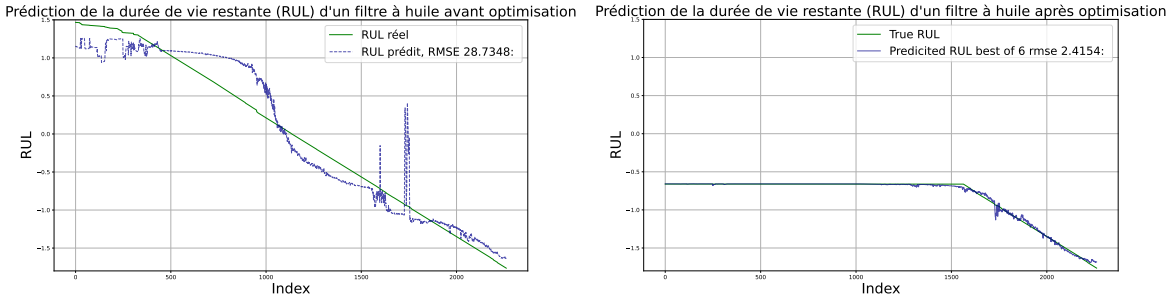
Cinq entraînements ont été réalisés pour chaque modèle. Comme précédemment un modèle moyen est considéré, dans ce cas là, la prédiction moyenne faite par les trois meilleurs modèles parmi les cinq.

Les figures 5a et 5b correspondent à l'estimation de la durée de vie restante et à la durée de vie restante réelle. Dans le cas de l'estimation directe du RUL avec DAST, ce sont les sorties du modèle qui sont affichés sur la figure 5a. Pour l'estimation du RUL avec DAST-OR, les sorties du modèle étant des vecteurs de probabilités, une transformation est faite pour obtenir un RUL comme dans Vishnu et al. (2019).

Les résultats dépendent beaucoup des données, de la répartition des filtres entres le jeu d'entraînement et de test. Les résultats obtenus avec DAST-OR avant optimisation sont peut-être liés au faible nombre de filtres à notre disposition. De meilleurs résultats pourraient être obtenus avec plus de données (comme les données terrains pour lesquels nous avons une quantité importante de données censurées). Les résultats de DAST sont prometteurs, une optimisation des hyperparamètres est en cours pour les deux modèles cf 6b.

Les prédictions figure 6 montrent les améliorations possibles en optimisant les paramètres des modèles. Un des paramètres clé concernant l'optimisation de DAST est le $RUL_{max}$. Comme

notre objectif est de prédire le RUL de chaque filtre, ce qui est crucial pour ce cas d'usage est d'être précis au sujet des prédictions effectuées à la fin de chaque durée de vie restante. En pratique, la dégradation du filtre au début de sa vie peut être considérée comme négligeable. Par conséquent, il est possible de fixer une valeur maximale pour la durée de vie restante ($RUL_{max}$) comme dans Li et al. (2020)) et Heimes (2008) sur données C-MAPSS. Pour DAST non optimisé, le $RUL_{max}$ n'a pas été fixé. L'étape suivante est d'optimiser DAST-OR, il devrait y avoir une amélioration significative des résultats comme celle observée sur DAST une fois optimisée. Lorsque les optimisations de DAST et DAST-OR seront faites, il sera alors possible de comparer leurs performances.




(a) Résultats de DAST avant optimisation sur le filtre à huile 2

(b) Résultats DAST après optimisation sur le filtre à huile 2

FIGURE 6 – Résultats de DAST et DAST-OR pour l'estimation directe du RUL du filtre à huile 1

# 4 Conclusion et perspectives

Les performances de DAST-OR sur les données C-MAPSS sont bonnes, c'est pourquoi une évaluation est effectuée sur l'application Liebherr. De plus, la proportion de filtres censurés sur les données Liebherr de terrains est très importante (>90%). D'où l'intérêt du modèle DAST-OR pour Liebherr, car il permet de prendre en compte la censure.

Après optimisation des modèles, des bons résultats sont obtenus sur les données Liebherr avec DAST. Une optimisation des paramètres sera aussi faite pour DAST-OR. Les paramètres optimisés sont les suivants, *le taux d'apprentissage, la taille de la fenêtre glissante, la taille des batchs, et la taille des matrices d'attention.* D'autres pistes sont à l'étude, comme le changement de fonction d'erreur en mettant une pondération inversement proportionnelle au RUL dans l'erreur (pour accorder plus d'importance aux prédictions faîtes à la fin de la durée de vie d'un filtre qu'au début) et une erreur asymétrique qui pénaliserait plus une sur-estimation qu'une sous-estimation du RUL cf Saxena et al. (2008).

# Bibliographie

Heimes, F. O. (2008). Recurrent neural networks for remaining useful life estimation. In *2008 international conference on prognostics and health management* (pp. 1–6).

Li, H., Zhao, W., Zhang, Y., & Zio, E. (2020). Remaining useful life prediction using multi-scale deep convolutional neural network. *Applied Soft Computing*, *89*, 106113.

Noot, J.-P., Rey, F., & Birmelé, E. (2024). Lstm and transformers based methods for remaining useful life prediction considering censored data. *article en cours de rédaction*.

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9).

Song, Y., Gao, S., Li, Y., Jia, L., Li, Q., & Pang, F. (2020). Distributed attention-based temporal convolutional network for remaining useful life prediction. *IEEE Internet of Things Journal*, *8*(12), 9594–9602.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vishnu, T., Malhotra, P., Vig, L., & Shroff, G. (2019). Data-driven prognostics with predictive uncertainty estimation using ensemble of deep ordinal regression models. *International Journal of Prognostics and Health Management*, *10*(4).

Zhang, Z., Song, W., & Li, Q. (2022). Dual aspect self-attention based on transformer for remaining useful life prediction. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–11.

# AnoRand: Deep Learning-Based Semi-Supervised Anomaly Detection with Synthetic Labels

Mansour Zoubeirou A Mayaki [1] & Michel Riveill [2]

[1] *Université Côte d'Azur, Inria, CNRS, Nice France, mansour.zoubeirou-a-mayaki@inria.fr*
[2] *Université Côte d'Azur, CNRS, Inria, Nice France, michel.riveill@unice.fr*

**Résumé.** Cet article présente AnoRand, une nouvelle méthode de détection d'anomalies semi-supervisée qui combine l'apprentissage profond avec la génération aléatoire de labels synthétiques. AnoRand comprend un bloc de détection de bruit (ND) et un bloc d'autoencodeur (AE). L'architecture tire parti de la capacité des modèles de autoencodeurs à représenter les données dans un espace latent et de celle des perceptrons à apprendre les caractéristiques d'une classe même en présence de données déséquilibrées. Pour enrichir l'ensemble d'apprentissage, nous générons des anomalies synthétiques en perturbant aléatoirement une petite portion des données normales. Cette technique permet d'élargir la diversité des exemples et d'améliorer la capacité du modèle à détecter des anomalies réelles. Nous avons comparé AnoRand à 17 autres méthodes de détection d'anomalies non supervisées sur des ensembles de données synthétiques et réels. Les résultats démontrent la performance supérieure d'AnoRand, qui obtient systématiquement les meilleurs scores en termes d'AUC ROC et d'AUC PR sur une large variété d'ensembles de données.
**Mots clés.** Détection d'anomalie, Réseaux de neurones, Apprentisage non supervisé.

**Abstract.** Anomaly detection, a challenging task in machine learning, often faces limited labeled data or none at all. This paper introduces AnoRand, a novel semi-supervised anomaly detection method blending deep learning with random synthetic label generation. AnoRand comprises a noise detection (ND) block and an autoencoder (AE) block. The architecture leverages autoencoders' latent space representation and feed forward perceptron's ability to handle imbalanced data. Synthetic anomalies are generated by randomly perturbing a small portion of the training set, then combined with normal samples for model input. Comparative evaluations against 17 unsupervised anomaly detection methods on synthetic and benchmark datasets demonstrate AnoRand's superior performance, consistently achieving the best results in terms of AUC ROC and AUC PR across diverse datasets.
**Keywords.** Anomaly detection, outlier detection, semi-supervised learning, deep learning, autoencoder,imbalanced data.

## 1 Introduction

Anomaly detection poses significant challenges in machine learning, particularly due to limited labeled samples and label accuracy issues. Unsupervised methods have gained popularity but are constrained by assumptions about anomaly distribution. Here, we introduce

AnoRand, a semi-supervised approach combining a deep autoencoder with feed forward perceptrons (FFP). AnoRand optimizes both models jointly, leveraging the autoencoder's latent space representation and FFP's ability to handle imbalanced data. In our method, the FFP block has a role in informing and strengthening the capacity of the auto-encoder block to embed the normal samples. We compared the performance of the proposed method to 17 state-of-the-art unsupervised anomaly detection method on synthetic data sets and 57 real-world data sets from the ADBench benchmark paper [1]. Our results show that this new method generally outperforms most of the state-of-the-art methods and has the best performance (AUC ROC and AUC PR) on the vast majority of reference datasets. In particular, AnoRand outperforms Deep auto encoder, Variational auto encoder and MLP even though they have the same kind of building blocks.

# 2 AnoRand method and implementation details

## 2.1 Proposed Architecture

Lets first define $(x, y) = \{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\}$ in the machine learning framework such that $y_i$ is the target (label) vector and $x_i \in R^d$ the feature vector for the $ith$ sample.

In this paper, we propose AnoRand, a novel method that combines an autoencoder with a fully connected architecture for anomaly detection. AnoRand optimizes a deep autoencoder and a Feed Forward Perceptron (FFP) model jointly to reduce reconstruction errors and enhance anomaly detection capabilities. While autoencoders excel at learning compact data representations, their effectiveness in detecting anomalies can be limited by imbalanced classes, resulting in blurred reconstructions that obscure subtle anomalies. AnoRand addresses this by incorporating the FFP's output into the autoencoder's latent vector, leveraging the FFP's ability to learn single classes in imbalanced data. The method involves two steps: synthetic anomaly generation to enrich the dataset and model input/training using both normal and synthetic anomalies. The FFP block, acting as a "noise detection block," is crucial for identifying anomalies, especially with noisy synthetic labels. We denote the mappings of input features by FFP and the encoder as $z_1 = F(x, \theta_0)$ and $z_0 = E(x, \theta_1)$ respectively, with $\theta_1$ and $\theta_0$ representing their respective weights. The final latent vector of the autoencoder block is defined as follows:

$$z = (z_0, z_1) = (F(x, \theta_0), E(x, \theta_1)) \tag{1}$$

The last layer of the FFP, denoted as $z_1 = F(x, \theta_0)$, produces an output based on the features it has learned. The combination of $z_0$ and $z_1$ allows for the integration of information captured by the autoencoder (data-specific features) with the features learned by the FFP (task-specific or high-level features). This combined vector is particularly powerful when we want to perform tasks like generating data samples that are consistent with both the learned data distribution and the task-specific features. The full network architecture is described in Figure 1.

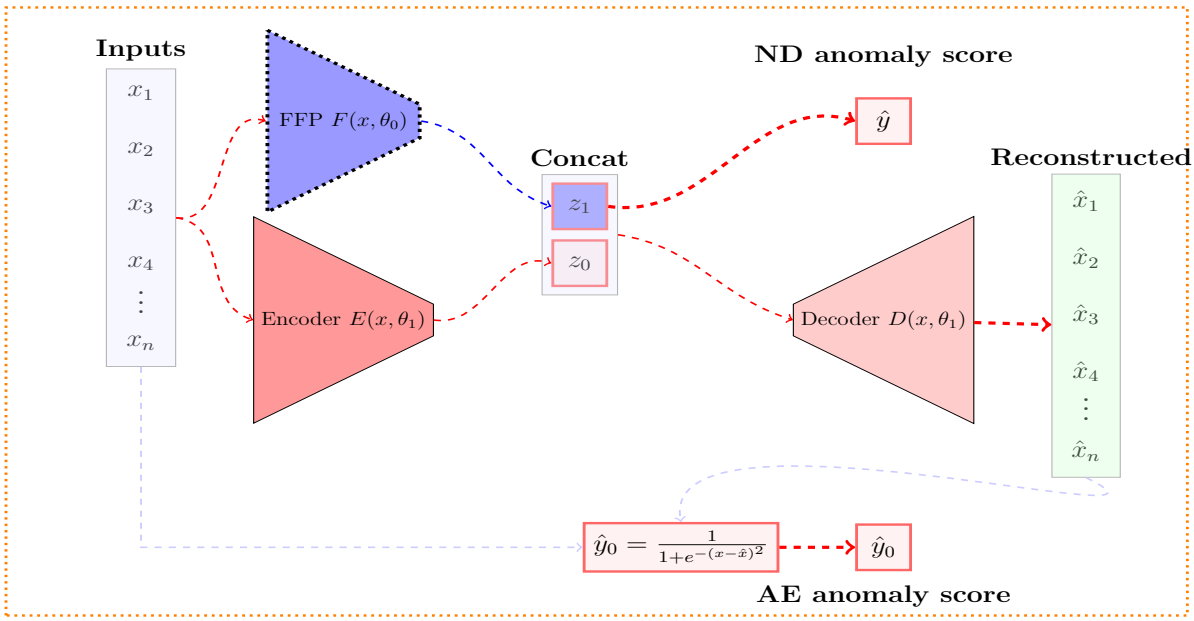



Figure 1: Overall architecture of AnoRand

## 2.2 Objective function

The final loss function $\mathcal{L}(\theta)$ combines two loss components with a weight factor $w$.
$\mathcal{L}(\theta_0)$: This loss is derived from the predictions of the noise detection (ND) block, evaluating the model's ability to accurately classify data points as either normal or anomalous. It also assesses how well the ND block can differentiate between regular data points and outliers or anomalies:

$$\mathcal{L}(\theta_0) = \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_0^i) = \sum_{i=1}^{N} y_i \log(\hat{y}_0^i) + (1 - y_i) \log(1 - \hat{y}_0^i) \tag{2}$$

$\mathcal{L}(\theta_1)$: This term represents the reconstruction error of the autoencoder (AE) block. It is calculated for each data point $\mathbf{x}_i$ and its corresponding reconstructed output $\hat{x}_i$. The goal of this component is to encourage the AE to learn meaningful representations of normal samples:

$$\mathcal{L}(\theta_1) = \sum_{i=1}^{N} \mathcal{L}(\mathbf{x}_i, \hat{x}_i) = \sum_{i=1}^{N} (1 - y_i)(\mathbf{x}_i - \hat{x}_i)^2 \tag{3}$$

The term $1 - y_i$ acts as a filter, focusing the loss calculation on normal samples ($y_i = 0$). The overall loss function combines these components, with the weight $w$ dictating their relative importance:

$$\mathcal{L}(\theta) = w \cdot \mathcal{L}(\theta_0) + (1 - w) \cdot \mathcal{L}(\theta_1) \tag{4}$$

Key aspects of this loss function include:

- $\theta = (\theta_0, \theta_1)$. $\theta_0$ and $\theta_1$ are respectively the parameters of the ND block and the AE block

and $\theta$ represents the overall model parameters. N is the total number of samples. $\hat{y}_0^i$ is the estimated probability that sample $i$ is an anomaly calculated by the ND block.

- $0 \leq w \leq 1$ balances the significance of the AE block's reconstruction capabilities against the ND block's noise detection capabilities within the loss function. A higher $w$ value emphasizes anomaly detection, while a lower value focuses on data reconstruction quality. This can be especially useful in situations where accurate anomaly identification is of paramount importance. In contrast, when $w$ is set lower, it places more emphasis on the autoencoder's ability to faithfully reconstruct the input data, which can be advantageous when the quality of the normal data reconstruction is a primary concern. The final loss can be rewritten as follows:

$$\begin{aligned}\mathcal{L}(\theta) = \mathcal{L}(\Phi(x), y) &= w \cdot \mathcal{L}(\theta_0) + (1-w) \cdot \mathcal{L}(\theta_1) = w \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_0^i) + (1-w) \sum_{i=1}^{N} \mathcal{L}(\mathbf{x}_i, \hat{x}_i) \\ &= \sum_{i=1}^{N} \left[ (1-w) \cdot (1-y_i)(\mathbf{x}_i - \hat{x}_i)^2 + w \left[ \cdot y_i \log(\hat{y}_0^i) + (1-y_i) \log(1-\hat{y}_0^i) \right] \right] \\ &= \sum_{i=1}^{N} \left[ w \cdot y_i \log(\hat{y}_0^i) + (1-y_i) \left[ (1-w) \cdot (\mathbf{x}_i - \hat{x}_i)^2 + w \cdot \log(1-\hat{y}_0^i) \right] \right] \end{aligned} \tag{5}$$

- $w \cdot y_i \log(\hat{y}_0^i)$: This term is associated with the noise detection (ND) block and focuses on the classification of data points as either anomalous ($y_i = 1$) or non-anomalous ($y_i = 0$). It calculates the logarithm of the predicted probability $\hat{y}_0^i$ for the true class labels $y_i$. When $y_i = 1$ (indicating an anomaly), this term encourages $\hat{y}_0^i$ to be close to 1 indicating high confidence in the anomaly prediction. It measures how well the model's predictions $\hat{y}_0^i$ align with the true labels $y_i$. When $y_i = 1$, this term evaluates how well the model predicts the probability of an anomaly ($\hat{y}_0^i$).

- $(1-y_i)\left[(1-w) \cdot (\mathbf{x}_i - \hat{x}_i)^2 + w \cdot \log(1-\hat{y}_0^i)\right]$: This part combines contributions from both the autoencoder (AE) block and the ND block. It quantifies how each block contributes into the loss of a negative sample.

  - $(1-w) \cdot (\mathbf{x}_i - \hat{x}_i)^2$: this component reflects the reconstruction error from the AE block. When $y_i = 0$ (non-anomaly), it encourages the squared difference between the input data $\mathbf{x}_i$ and its reconstruction $\hat{x}_i$ to be minimized. This term drives the AE to capture meaningful data representations. When $w$ is closer to 1, this term contributes less to the loss, emphasizing data reconstruction.
  - $w \cdot \log(1-\hat{y}_0^i)$: This component is related to the anomaly detection objective. It encourages the logarithm of $(1-\hat{y}_0^i)$ when $y_i = 0$. In other words, it encourages the model to assign lower probabilities to non-anomalous data points. It assesses how well the model predicts the probability of a non-anomaly (1 minus the probability of an anomaly, $1-\hat{y}_0^i$) when $y_i = 0$ (indicating a negative class or non-anomaly).

- **Reconstruction Quality:** Encouraged by the $(1-w) \cdot (\mathbf{x}_i - \hat{x}_i)^2$ term, this part of the loss function motivates the AE block to learn meaningful data representations. It measures how

accurately the model can reconstruct input data when the data is non-anomalous ($y_i = 0$). A lower reconstruction error implies that the AE is successful in capturing essential features of the data.

- **Anomaly Detection**: Guided by the $w \cdot y_i \log(\hat{y}_0^i)$ and $w \cdot \log(1 - \hat{y}_0^i)$ terms, the ND block focuses on accurately classifying data points as anomalies or non-anomalies. It encourages the model to assign high probabilities ($\hat{y}_0^i$ close to 1) to anomalies ($y_i = 1$) and low probabilities ($\hat{y}_0^i$ close to 0) to non-anomalies ($y_i = 0$).

The objective during model training is to find the optimal configuration of model parameters $\hat{\theta}$ that minimizes the training loss $\mathcal{L}(\theta)$. The optimization objective is defined as:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta) = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(\Phi(x_i, \theta), y_i) \tag{6}$$

### 2.3 Anomaly score

In the AnoRand model, we use two outputs: $\hat{y}_0$ from the ND block and $\hat{x}$ from the AE block's reconstruction, for classifying input samples as anomalies or normal data. We combine these outputs to calculate the final anomaly score. We define $\hat{y}_1$ as the predicted probability of a sample being an anomaly based on the AE block reconstruction $\hat{x}$. Higher reconstruction errors indicate a higher likelihood of being an anomaly. To balance $\hat{y}_1$ and $\hat{y}_0$, we introduce a weight parameter $\alpha$, which adapts the model's behavior based on the dataset and the noise detection versus reconstruction trade-off. $\alpha$ is calculated from the third quantile of the anomaly score from both blocks. Specifically, $\alpha$ and $\hat{y}_0$ are determined as follows:

$$\hat{y}_1 = \frac{1}{1 + e^{-(x-\hat{x})^2}} \qquad\qquad \alpha = \frac{Q_3^1}{Q_3^0 + Q_3^1}$$

$Q_3^1$ represents the third quantile of the autoencoder block's predicted probabilities $\hat{y}_1$. $Q_3^0$ represents the third quantile of the noise detection block's predicted probabilities $\hat{y}_0$.
Using quantiles, such as the third quantile, provides a robust measure for estimating the range of predicted probabilities. This approach is less sensitive to extreme values and outliers, making it suitable for ensuring that the weight $\alpha$ is derived from a reliable range of values. With the weight $\alpha$ determined, the final anomaly score denoted as $\hat{y}_{score}$ is computed as a weighted sum of $\hat{y}_0$ and $\hat{y}_1$:

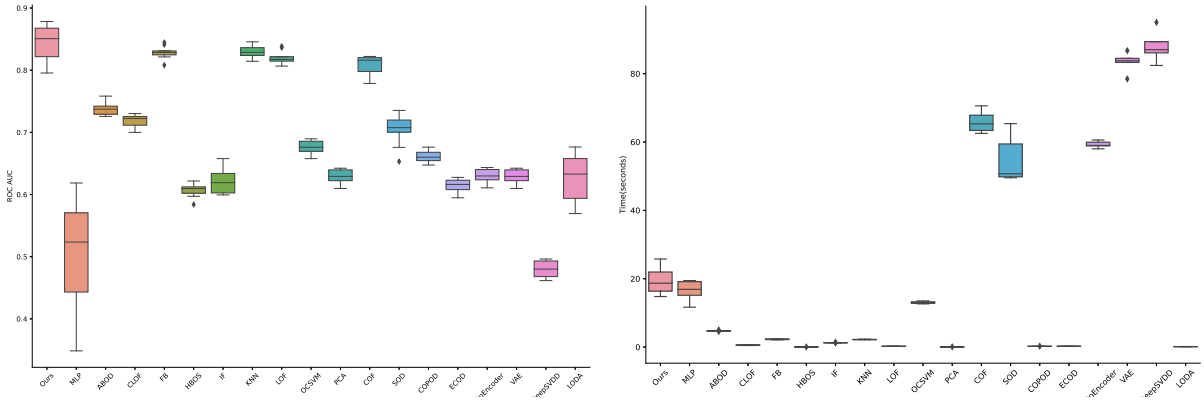
$$\hat{y}_{score} = \hat{y}_1 \cdot (1 - \alpha) + \alpha \cdot \hat{y}_0 \tag{7}$$

## 3 Experiment

### 3.1 Comparative Study of Anomaly Detection Algorithms

Figure 2 shows that our proposed method consistently outperformed all other algorithms in terms of AUC PR and AUC ROC. When compared to deep learning-based unsupervised

methods like Deep Autoencoder (AE), Variational Autoencoder (VAE), and Multi-Layer Perceptron (MLP), our approach demonstrated superior performance, despite sharing similar architectural foundations. Intriguingly, deep learning-based unsupervised methods, such as DeepSVDD and Autoencoder, exhibited unexpectedly subpar performance compared to classical techniques. It is worth noting that while our method achieved good performance, it came at the cost of increased training time. In Figure 2b, we provide a visual representation of the comparison of training duration between the algorithms. The reasons behind this prolonged training time may be attributed to the complexity of our model, the number of training iterations and the number of parameters.




(a) AUC PR (b) Training time in seconds

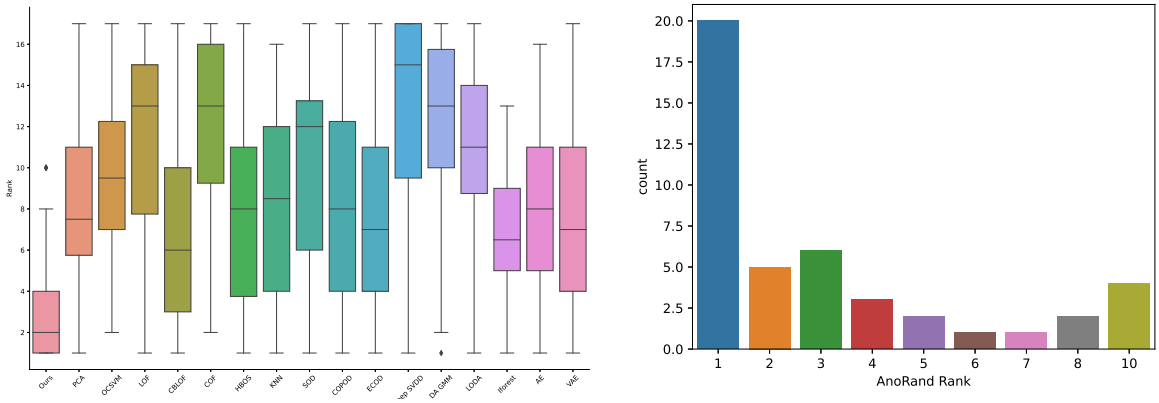
Figure 2: Performance metrics on synthetic data set for unsupervised algorithms

## 3.2 Results and discussions on benchmark

We conducted a comprehensive comparison of the performance of our method (**AnoRand**) against state-of-the-art unsupervised methods using the ADBench anomaly detection benchmark [1]. The benchmark, as described in their paper, evaluates the performance of 17 algorithms on 57 benchmark datasets, and for our experiments, we selected 44 of the most challenging datasets for comparison. These datasets span various domains, including healthcare, security, and more. We categorized the datasets into four groups to facilitate comparison: Image datasets, NLP datasets, Healthcare datasets and Science datasets, along with datasets from other fields such as documents and the web. Detailed results and performances from all experiments are thoroughly documented in Table 1. In Figure 3, we reported the algorithms performance and their rankings on the ADBench benchmark datasets. The experimental results also show that our model has the best overall ranking among all its counterpart unsupervised algorithms. In terms of AUC PR, AnoRand is ranked first ($1^{st}$) on 20 datasets, second ($2^{nd}$) on 5, third ($3^{rd}$) on 6 and fourth ($4^{th}$) on 3. The results also show that, in situations where another algorithm outperforms ours, the performance gap is very small in most cases.

Table 1: AUCPR (in %) of 16 unsupervised algorithms on 44 benchmark datasets. For our method, we computed the average AUCPR over 10 runs and added the standard deviation. The performance rank is shown in parentheses (the lower, the better), and mark the best performing method(s) in **bold**.

| Category | Dataset | PCA | OCSVM | LOF | CBLOF | COF | HBOS | KNN | SOD | COPOD | ECOD | Deep SVDD | DA GMM | LODA | Iforest | AutoEncoder | VAE | **AnoRand(Ours)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Image and CV | mnist | 39,93(3) | 33,2(6) | 20,9(14) | 28,82(8) | 25,51(11) | 12,51(17) | 35,53(5) | 19,15(16) | 21,35(13) | 31,93(7) | 19,72(15) | 23,75(12) | 25,86(10) | 27,71(9) | 40,28(2) | 39,87(4) | **50.17** ±1.25 **(1)** |
| | optdigits | 2,76(14) | 2,92(13) | 6,06(4) | 10,08(2) | 4,42(7) | 10,03(3) | 3,06(12) | 4,39(8) | 4,36(9) | 3,43(11) | - | 5,59(5) | 3,95(10) | 5,09(6) | 2,64(16) | 2,67(15) | **56.34** ±5.42**(1)** |
| | skin | 17,4(14) | 19,03(9) | 18,25(12) | 29,82(2) | 16,38(15) | 23,7(8) | 28,72(4) | 24,61(7) | 17,99(13) | 15,96(16) | 18,48(10) | - | 18,44(11) | 26,08(6) | 28,08(5) | 29,78(3) | **52.20** ±12.26**(1)** |
| | FashionMNIST | 31,42(9) | 31,97(8) | 16,85(16) | 38,9(2) | 20,73(14) | 29,43(11) | 33,87(3) | 28,72(12) | 30,32(10) | 32,53(5) | 17,43(15) | 14,44(17) | 27,32(13) | 32,35(7) | 32,46(6) | 32,66(4) | **43.16** ±2.48**(1)** |
| | MNIST-C | 16,88(10) | 17,72(7) | 13,84(15) | 27,62(2) | 14,53(14) | 15,46(13) | 22,98(3) | 15,68(12) | 15,9(11) | 18,24(5) | 8,34(17) | 11,37(16) | 18,63(4) | 17,99(6) | 17,03(9) | 17,24(8) | **35.21** ±1.26**(1)** |
| | satimage-2 | 85,69(6) | 82,71(7) | 4,29(16) | 97,09(1) | 8,81(15) | 78,04(9) | 39,14(12) | 26,11(13) | 76,55(10) | 63,25(11) | 3,08(17) | 22,07(14) | 80,52(8) | 93,45(3) | 86,36(5) | 87,52(4) | 94.12 ±01.80(2) |
| | MVTec-AD | 54,06(9) | 51,44(13) | 54,9(7) | 58,52(1) | 46,59(15) | 55,22(6) | 55,55(4) | 51,48(12) | 54,64(8) | 55,44(5) | 36,5(17) | 45,66(16) | 49,73(14) | 56,04(3) | 51,57(11) | 51,62(10) | 57.65±0.12(2) |
| | letter | 6,86(15) | 6,11(17) | 34,02(1) | 14,8(6) | 21,43(5) | 8,38(10) | 30(2) | 28,63(3) | 6,77(16) | 6,94(13) | 9,29(8) | 11,68(7) | 6,87(14) | 8,49(9) | 8,25(12) | 8,31(11) | 28.47 ±1.84(4) |
| | celeba | 15,89(1) | 10,73(8) | 1,71(17) | 11,33(7) | 1,77(16) | 13,82(2) | 3,14(12) | 2,66(13) | 13,69(3) | 12,37(4) | 2,34(14) | 1,95(15) | 4,04(11) | 8,96(9) | 11,39(5) | 11,39(5) | 5.61 ±0.50(10) |
| | CIFAR10 | 10,59(6) | 10,19(9) | 13,02(1) | 10,61(5) | 11,61(2) | 8,38(15) | 11,13(3) | 11,06(4) | 8,77(14) | 9,29(12) | 8,05(16) | 7,73(17) | 9,72(11) | 8,97(13) | 10,32(8) | 10,45(7) | 10.04 ±0.91 (10) |
| NLP | speech | 1,97(13) | 1,96(14) | 2,52(3) | 1,99(12) | 2,25(5) | 2,09(9) | 2,02(11) | 2,13(8) | 1,94(15) | 1,77(17) | 5,12(2) | 2,03(10) | 1,79(16) | 2,31(4) | 2,16(6) | 2,16(6) | **7.65** ±0.11 **(1)** |
| | Imdb | 4,55(15) | 4,44(17) | 4,83(8) | 4,75(9) | 5,16(4) | 4,74(10) | 4,49(16) | 4,7(12) | 4,9(6) | 4,9(6) | 5,06(5) | 4,65(13) | 4,59(14) | 4,74(10) | 6,18(3) | 6,3(2) | **7.65**±0.25 **(1)** |
| | Agnews | 5,74(11) | 5,69(12) | 14,35(1) | 7,02(6) | 12,21(2) | 5,58(13) | 8,61(4) | 8,4(5) | 5,43(14) | 5,43(14) | 4,45(17) | 5,41(16) | 5,93(10) | 6,04(9) | 6,18(8) | 6,3(7) | 9.04 ±1.58 (3) |
| | Amazon | 5,85(10) | 5,64(16) | 5,72(13) | 6,07(5) | 5,74(12) | 5,98(7) | 6,23(2) | 6,4(1) | 6,08(4) | 6,06(6) | 3,84(17) | 5,65(15) | 5,92(9) | 5,95(8) | 5,75(11) | 5,69(14) | 6.20 ±0.23 (3) |
| | Yelp | 7,62(13) | 7,75(10) | 8,52(5) | 7,68(11) | 8,68(4) | 7,81(9) | 9,85(1) | 9,2(2) | 8,01(6) | 7,98(7) | 6,39(17) | 6,72(16) | 7,65(12) | 7,88(8) | 7,05(15) | 7,14(14) | 8.80 ±1.32 (3) |
| Healthcare | WBC | 82,29(9) | 89,87(6) | 5,57(16) | 92,27(4) | 9,73(14) | 73,56(11) | 66,55(12) | 54(13) | 86,19(7) | 86,19(7) | 6,38(15) | #N/A | 78,67(10) | 90,49(5) | 94,3(3) | 94,45(2) | **94.55** ±4.66**(1)** |
| | Cardiotocography | 47,95(6) | 52,61(2) | 30,66(14) | 45,44(7) | 28,21(16) | 38,28(11) | 34,79(12) | 27,99(17) | 40,46(10) | 43,57(8) | 34,03(13) | 30,61(15) | 48(5) | 41,47(9) | 48,49(3) | 48,16(4) | **61.39** ±5.74**(1)** |
| | Lymphography | 97,02(6) | 93,59(7) | 23,08(14) | 97,62(2) | 36,68(13) | 91,83(8) | 38,69(12) | 22,65(15) | 88,68(10) | 90,87(9) | 4,58(17) | 19,52(16) | 44,54(11) | 97,31(3) | 97,22(5) | 97,24(4) | **99.68** ±0.01**(1)** |
| | breastw | 95,11(7) | 82,7(13) | 28,55(15) | 91,54(10) | 27,6(16) | 97,71(4) | 92,19(9) | 84,88(12) | 99,4(1) | 98,54(2) | 50,92(14) | - | 97,04(5) | 96,04(6) | 89,17(11) | 94,96(8) | 98.17 ±0.40(3) |
| | WPBC | 23,01(8) | 22,93(9) | 20,29(17) | 21,32(15) | 21,3(16) | 23,04(7) | 21,49(13) | 25,37(3) | 22,81(10) | 21,38(14) | 26,24(1) | 22,49(11) | 25,58(2) | 22,42(12) | 23,91(4) | 23,44(5) | 23.06 ±0.44(6) |
| | Hepatitis | 36,65(4) | 29,44(10) | 13,69(17) | 31,54(8) | 14,39(16) | 37,73(3) | 21,95(15) | 24,89(12) | **41.5(1)** | 37,82(2) | 22,17(14) | 22,96(13) | 30,9(9) | 26,25(11) | 33,81(6) | 33,58(7) | 35.20 ±7.17(5) |
| | thyroid | 44,34(6) | 21,23(12) | 20,81(13) | 29,95(9) | 28,5(10) | 50,98(3) | 34,98(8) | 23,56(11) | 19,64(14) | 54,05(2) | 2,5(17) | 16,06(15) | 14,68(16) | **63.11(1)** | 46,16(5) | 36,9(7) | 47.79 ±2.77(4) |
| | annthyroid | 16,12(11) | 10,37(15) | 15,71(12) | 13,69(14) | 14,39(13) | 16,99(8) | 16,74(9) | 18,84(6) | 16,58(10) | 24,65(2) | 21,95(4) | 9,64(16) | 7,06(17) | **30.47(1)** | 22,73(3) | 19,4(5) | 17.02 ±0.80 (5) |
| | Pima | 54,03(5) | 50(8) | 47,18(11) | 53,19(6) | 44,7(13) | **56.61(1)** | 55,14(4) | 48,24(10) | 55,19(3) | 37,3(16) | 35,87(17) | 41,55(15) | 44,09(14) | 55,82(2) | 46,56(12) | 49,53(9) | 50.09±0.89(7) |
| | cardio | 66,06(2) | 62,89(5) | 23,79(16) | 61,95(6) | 28,67(14) | 52,1(11) | 40,72(12) | 28,54(15) | 60,42(7) | **68.59(1)** | 22,5(17) | 28,92(13) | 53,41(9) | 59,95(8) | 64,72(3) | 62,9(4) | 52.53 ±4.68 (8) |
| Others | musk | 99,89(4) | 10,61(11) | 2,82(15) | 100(1) | 2,61(16) | 100(1) | 9,65(12) | 7,59(13) | 34,79(9) | 34,95(8) | 5,39(14) | 32,75(10) | 47(7) | 99,61(6) | 100(1) | 99,89(4) | **100** ±0**(1)** |
| | Waveform | 5,79(12) | 4,37(16) | 11,33(5) | 18,98(2) | 14,11(3) | 5,86(11) | 13,04(4) | 9,6(6) | 6,9(7) | 6,8(8) | 4,83(13) | 3,11(17) | 4,71(14) | 6,2(9) | 6,2(9) | 4,65(15) | **33.26** ±0.79**(1)** |
| | cover | 9,8(7) | 11,41(5) | 8,1(9) | 5,83(15) | 4(16) | 6,83(13) | 6,16(14) | 3,88(17) | 11,37(6) | 15,63(3) | 8,1(9) | 27,59(2) | 13,06(4) | 8,8(8) | 7,27(11) | 7,25(12) | **34.19** ±0.54**(1)** |
| | fault | 32,76(14) | 38,44(8) | 38,38(9) | 43,98(4) | 41,56(5) | 36,47(10) | 54,45(2) | 48,01(3) | 30,54(17) | 30,82(16) | 39,15(7) | 33,48(13) | 31,03(15) | 41,09(6) | 34,58(11) | 34,46(12) | **63.20** ±1.51 **(1)** |
| | donors | 17(6) | 9,8(11) | 7,88(14) | 6,89(15) | 8,8(13) | 23,36(2) | 14,75(7) | 9,69(12) | 21,58(4) | 14,17(8) | 6,38(16) | 10,53(10) | 3,78(17) | 12,74(9) | 22,74(3) | 18,78(5) | **90.85** ±6.83**(1)** |
| | PageBlocks | 51,71(5) | 49,14(8) | 39,64(13) | 49,65(7) | 41,02(12) | 33,32(16) | 45,39(11) | 37,83(14) | 37,65(15) | 49(9) | 31,45(17) | 53,25(3) | 51,29(6) | 46,04(10) | 59,18(2) | 51,96(4) | **65.26** ±11.50**(1)** |
| | magic.gamma | 59,27(7) | 51,43(15) | 54,76(12) | 68,85(3) | 54,12(14) | 62,41(6) | 75,63(2) | 67,89(4) | 59,18(8) | 54,38(13) | 49,17(16) | 46,92(17) | 58,49(11) | 64,72(5) | 59,18(8) | 59,11(10) | **77.93** ±1.29**(1)** |
| | fraud | 22,91(11) | 47,58(2) | 47(4) | 47,52(3) | 22,86(12) | 25,89(10) | 47(4) | 31,37(9) | 42,82(8) | 42,99(7) | 8,97(17) | 21,32(14) | 46,37(6) | 21,67(13) | 15,88(15) | 15,88(15) | **60.08** ±0.15 **(1)** |
| | vertebral | 10,49(11) | 10,94(9) | 14,24(3) | 11,58(6) | 13,85(4) | 9,23(16) | 10,57(10) | 11,79(5) | 8,89(17) | 11,24(7) | 10,49(11) | 15,24(2) | 9,68(15) | 10,46(13) | 11,18(8) | 9,85(14) | **20.47** ±0.11**(1)** |
| | SpamBase | 41,57(7) | 40,12(12) | 35,16(15) | 41,18(11) | 34,73(16) | 50,03(5) | 41,42(8) | 40,03(13) | **56.68(1)** | 53,95(3) | 42,23(6) | - | 35,88(14) | 51,75(4) | 41,21(10) | 41,42(8) | 55.17 ±0.50(2) |
| | landsat | 16,18(17) | 16,21(16) | 24,69(7) | 30,97(3) | 24,95(6) | 22,03(10) | 24,65(8) | 26,38(4) | 17,48(14) | 25,17(5) | **38.83(1)** | 24,48(9) | 18,86(13) | 19,81(12) | 16,75(15) | 20,58(11) | 38.54 ±4.54 (2) |
| | shuttle | 92,35(7) | 85,29(10) | 13,76(16) | 60,98(11) | 12,17(17) | 96(4) | 20,38(13) | 20,27(14) | 96,56(2) | 95,76(5) | 15,86(15) | 93,2(6) | 48,75(12) | 91,27(9) | 91,49(8) | **97.62(1)** | 96.52 ±0.05(3) |
| | Stamps | 41,09(4) | 31,39(11) | 21,29(14) | 23,66(12) | 16,5(16) | 35,24(9) | 23,53(13) | 20,28(15) | 43(2) | 38,17(7) | 11,4(17) | **43.72(1)** | 34(10) | 39,49(5) | 38,35(6) | 38,11(8) | 41.66 ±0.39(3) |
| | satellite | 59,64(9) | 57,61(11) | 37,68(17) | 61,48(8) | 39,7(16) | 67,25(2) | 50,01(13) | 47,23(14) | 56,58(12) | 65,94(4) | 40,11(15) | 58,33(10) | 61,94(6) | 65,92(5) | 61,7(7) | 67,24(3) | **73.24** ±0.49**(1)** |
| | wine | 30,87(5) | 21,56(8) | 7,77(16) | 5,83(17) | 8,45(13) | 43,08(3) | 8,43(14) | 7,95(15) | 45,71(2) | 18,37(11) | 21,14(9) | 17,51(12) | **48.82(1)** | 25,96(6) | 19,1(10) | 22,02(7) | 41.59 ±1.76(4) |
| | campaign | 27(9) | 29,22(5) | 14,51(14) | 23,99(10) | 13,01(16) | 37,99(2) | 27,18(8) | 18,88(12) | **38.58(1)** | 37(3) | 11,6(17) | 14,62(13) | 13,47(15) | 32,26(4) | 28,51(6) | 28,51(6) | 23.99 ±1.76 (10) |
| | http | 56,43(3) | 46,86(5) | 3,82(14) | 47,53(4) | 9,57(12) | 44,79(6) | 0,7(15) | 8,32(13) | 35,19(7) | 16,61(11) | 29(8) | | 0,67(16) | **90.83(1)** | 25,49(10) | 25,55(9) | 60.24 ±0.07(2) |
| | InternetAds | 32,55(11) | 54,68(2) | 40,49(9) | **58.13(1)** | 38,67(10) | 53,97(3) | 43,23(7) | 27,69(15) | 50,97(5) | 51,07(4) | 27,91(12) | - | 23,89(16) | 48(6) | 27,72(14) | 27,73(13) | 42.79 ±1.43 (8) |
| | census | **10.02(1)** | 6,76(12) | 5,45(13) | 7,4(9) | 4,88(16) | 8,6(6) | 9(4) | 8,5(7) | 9,9(2) | 9,7(3) | 6,87(11) | 8,7(5) | 5,01(15) | 7,7(8) | 5,41(14) | 4,45(17) | 7.31 ±0.10 (10) |




(a) Overall AUC PR ranks (b) AnoRand AUC PR overall ranks

Figure 3: Algorithms performance rankings on real-world data sets (lower the better).

# 4 Conclusion

In this paper, we proposed a new semi supervised anomaly detection method based on deep autoencoder architecture. This new method that we called **AnoRand**, jointly optimizes the deep autoencoder and the FFP model in an end-to-end neural network fashion. Our

method is performed in two steps: we first create synthetic anomalies by randomly adding noise to few samples from the training data; secondly, we train our deep learning model in a supervised way with the new labeled data. Our method takes advantage of these limitations of FFP models in case of imbalance classes and use them to reinforce the autoencoder capabilities. Our experimental results show that our method achieves state-of-the-art performance on synthetic datasets and 57 benchmark datasets, significantly outperforming existing unsupervised alternatives. Moreover, on most benchmark datasets, regardless of the category, AnoRand outperforms all its deep learning-based counterparts. The main limitation of our method is that the training time is longer compared to most state-of-the-art non-deep learning algorithms, although it remains shorter than that of deep learning algorithms.

# References


[1] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[5] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

[6] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019.

[7] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020.

[8] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2041–2050, 2018.

[9] Guansong Pang, Chunhua Shen, Huidong Jin, and Anton van den Hengel. Deep weakly-supervised anomaly detection. *arXiv preprint arXiv:1910.13601*, 2019.

[10] Yingjie Zhou, Xucheng Song, Yanru Zhang, Fanxing Liu, Ce Zhu, and Lingqiao Liu. Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2454–2465, 2021.

[11] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[12] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 353–362, 2019.

# Modèle de réseau de neurones fiable pour accélérer les simulations couplées de thermodiffusion

M.B Yahiaoui[1] & L. Giraldi[1] & G. Daniel[2] & C. Introïni[1] & J. Arbel[3]

[1] *CEA, DES, IRESNE, DEC, SESC, F-13108 Saint-Paul-lez-Durance, France*
*{mohamed-bahi.yahiaoui, loic.giraldi, clement.introini}@cea.fr*

[2] *Université Paris-Saclay, CEA, SGLS, 91191, Gif-sur-Yvette, France*
*geoffrey.daniel@cea.fr*

[3] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
*julyan.arbel@inria.fr*

**Résumé.** Nous présentons une implémentation d'un modèle de substitution d'apprentissage profond pour prédire les contributions thermodynamiques (potentiels chimiques, mobilités) impliquées dans la résolution d'un problème d'inter-diffusion. Cette approche vise à accélérer les simulations, en tirant parti de la puissance des réseaux de neurones profonds. Le modèle de substitution est conçu pour reproduire le comportement du code de calcul d'équilibre thermodynamique, ici OpenCalphad, dans le cas simplifié d'un problème d'inter-diffusion pour un système ternaire. Les modèles d'apprentissage profond ont montré des performances remarquables sur notre cas d'étude sur nos données d'entraînement et de tests. Toutefois, il est nécessaire d'apporter des garanties sur l'erreur potentiellement commise par notre modèle. Nous abordons ainsi un aspect important dans l'utilisation des réseaux de neurones comme modèle de substitution et de leur fiabilité : les incertitudes. Nous présentons ainsi une méthode novatrice fondée sur l'étude du comportement des couches cachées des réseaux de neurones et visant à gérer efficacement l'incertitude dans l'apprentissage profond.

**Mots-clés.** Modèles de substitution, Apprentissage profond, Réseaux de neurones, Incertitudes, Surveillance en temps réel, Détection des données hors distribution, OpenCalphad, Thermodiffusion.

**Abstract.** We present an implementation of a deep learning surrogate model to predict the thermodynamic contributions (chemical potentials, mobilities) involved in solving an inter-diffusion problem. This approach aims to accelerate thermodiffusion simulations by leveraging the power of deep neural networks. The surrogate model is designed to replicate the behavior of the thermodynamic equilibrium calculation code, here OpenCalphad, in the simplified case of an inter-diffusion problem for a ternary system. Deep learning models have shown remarkable performance on our case study with our training and testing data. However, it is necessary to provide guarantees on the potential error made by our model. We thus address an important aspect in the use of neural networks as surrogate models and their reliability : uncertainties. We present an innovative method based on the study of the behavior of the hidden layers of neural networks and aiming to effectively manage uncertainty in deep learning.

**Keywords.** Surrogate models, Deep learning, Neural networks, Uncertainties, Runtime monitoring, Out-Of-Distribution detection, OpenCalphad, Thermodifusion.

# 1 Introduction

Dans le contexte de la modélisation de systèmes complexes, la construction de modèles précis et efficaces est essentielle pour comprendre et prédire le comportement de ces systèmes. Cependant, la complexité de certains problèmes, par exemple en modélisation physique, rend souvent la simulation directe difficile. C'est ici que les modèles de substitution, parfois désignés sous le terme de métamodèles, prennent leur place.

Les modèles de substitution [1, 9] sont des approximations simplifiées du système d'origine pour remplacer efficacement des simulations coûteuses en temps et en ressources de calcul. Ces modèles sont généralement construits par des méthodes d'apprentissage via des données obtenues à partir de simulations du système d'origine ou d'observations expérimentales.

Dans ce travail, l'objectif est d'utiliser un modèle de substitution pour accélérer la simulation de l'évolution d'un système ternaire de diffusion d'éléments chimiques dans un solide sous l'effet d'un gradient de température (ici, un profil parabolique imposé). Dans le système d'équations aux dérivées partielles associé à ce problème, les flux de diffusion dépendent des mobilités et des potentiels chimiques des éléments. Ces contributions sont évaluées à chaque pas de temps, en chaque point du maillage, par un code de calcul d'équilibres thermodynamiques, OpenCalphad [8]. De manière générale, ce type de couplage entre un code de calcul thermodynamique et un modèle de transport diffusif peut s'avérer coûteux selon la complexité de la description thermodynamique (non décrite dans ce travail) et la raideur du problème numérique. Dans ces conditions, l'utilisation d'un modèle de substitution est pertinente pour réduire le surcoût des simulations provoqué par la détermination des potentiels chimiques et des mobilités par les calculs thermodynamiques locaux effectués par OpenCalphad. De plus, l'étude d'un système simplifié permet de mettre en place une démarche générale et reproductible pour des études plus complexes impliquant des systèmes multiphasiques multi-composants [4, 5].

Les réseaux de neurones sont des outils intéressants pour construire des modèles de substitution. L'un des principaux avantages de l'utilisation des réseaux de neurones dans ce cadre est leur capacité d'approximation universelle [6]. Par conséquent, ils peuvent gérer des relations complexes qui peuvent exister dans le système sous-jacent, sous réserve d'avoir un nombre de données suffisant. L'une des limites de l'utilisation des réseaux de neurones est leur incapacité à fournir nativement une incertitude liée à leur manque de connaissance ou d'information, couramment appelée incertitude épistémique. Le principal défi est la dépendance du modèle à la distribution des données d'entraînement. Lorsque le réseau de neurones est déployé dans un environnement de production, il est nécessaire de s'assurer que les nouvelles évaluations du métamodèle soient valides, surtout pour les entrées qui s'éloignent de la distribution d'entraînement. Par conséquent, il est nécessaire de disposer d'un domaine opérationnel bien défini qui puisse être utilisé pour identifier les potentielles erreurs de prédiction avant qu'elles ne se propagent à travers le système. Pour garantir sa fiabilité en déploiement, le modèle doit être capable de détecter des données pour lesquelles il fournit des prédictions potentiellement erronées, appelées données hors domaine du modèle, ou *Out-of-Model Scope (OMS) data* [2]. À cette fin, nous introduisons une nouvelle approche de detection des données hors domaine du modèle, nommée **BB-AS** (*Bounding Box with*

*Anomaly Score*). Elle consiste à comparer les états des couches intermédiaires avec un ensemble de boîtes englobantes [3].

Dans la section 2, nous décrivons brièvement le problème d'inter-diffusion pour un système ternaire et le développement d'un métamodèle pour le calcul des contributions thermodynamiques. Puis, dans la section 3, nous présentons la méthode BB-AS en décrivant les principes de son fonctionnement et son application au modèle de substitution d'OpenCalphad.

# 2 Métamodélisation du code de calcul OpenCalphad

## 2.1 Description du problème physique

Dans le cadre de ce travail, nous nous intéressons à la simulation de l'évolution d'un système ternaire U-Pu-O en phase solide. Plus spécifiquement, ces travaux visent à accélérer les simulations numériques de diffusion d'espèces, où les forces motrices sont exprimées en fonction des potentiels chimiques calculés par OpenCalphad.

La simulation nécessite d'abord la définition d'un état initial, comprenant les profils des quantités molaires initiales $N_e^0(.)$ des éléments dans le système ainsi que le profil de température $T(.)$ imposé. Les calculs d'équilibres thermodynamiques sont réalisés à une pression constante de 50 bars. Puis, l'algorithme de résolution est basé sur un schéma partitionné en deux étapes, illustré sur la Figure 1.

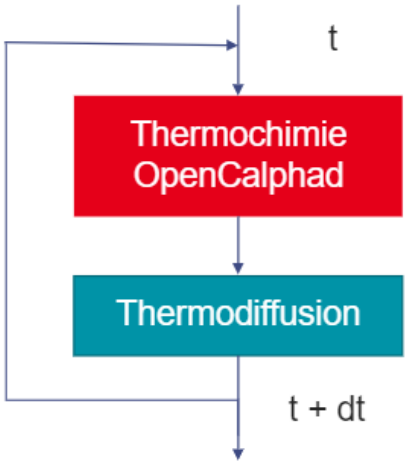



FIGURE 1 – Algorithme partitionné pour la résolution du problème d'inter-diffusion couplé au code de calcul thermodynamique OpenCalphad.

À chaque pas de temps et sur chaque nœud spatial du système, la première étape consiste à réaliser des calculs d'équilibres thermodynamiques OpenCalphad pour déterminer les potentiels chimiques $\mu_e$ et les mobilités $M_e$ de chaque élément $e$ (Figure 2). Ensuite, ces


FIGURE 2 – Entrées et sorties du composant OpenCalphad, que l'on cherche à métamodéliser

données thermodynamiques sont utilisées à la deuxième étape pour résoudre les équations d'interdiffusion suivantes (Éqs. 1) :

$$\frac{\partial x_{\mathrm{O}}}{\partial t} = \nabla \cdot [M_{\mathrm{O,U}} \nabla(\mu_O - \mu_{\mathrm{U}}) + M_{\mathrm{O,Pu}} \nabla(\mu_{\mathrm{O}} - \mu_{\mathrm{Pu}})], \tag{1a}$$

$$\frac{\partial x_{\mathrm{U}}}{\partial t} = \nabla \cdot [M_{\mathrm{O,U}} \nabla(\mu_{\mathrm{U}} - \mu_{\mathrm{O}}) + M_{\mathrm{U,Pu}} \nabla(\mu_{\mathrm{U}} - \mu_{\mathrm{Pu}})], \tag{1b}$$

$$\frac{\partial x_{\mathrm{Pu}}}{\partial t} = \nabla \cdot [M_{\mathrm{O,Pu}} \nabla(\mu_{\mathrm{Pu}} - \mu_{\mathrm{O}}) + M_{\mathrm{U,Pu}} \nabla(\mu_{\mathrm{Pu}} - \mu_{\mathrm{U}})], \tag{1c}$$

où $x_e$ est la fraction molaire de l'élément $e$, calculée à partir des quantités molaires des éléments, $M_{e,e'}$ le coefficient d'inter-diffusion entre les éléments $e$ et $e'$. Le coefficient d'inter-diffusion est calculé à partir des mobilités $M_e$, $M_{e'}$ et des fractions molaires $x_e$, $x_{e'}$ de chaque élément :

$$M_{e,e'} = (M_e x_{e'} + M_{e'} x_e) x_e x_{e'}.$$

A l'issue de la deuxième étape, les quantités molaires de chaque élément sont mises à jour pour les calculs d'équilibres thermodynamique de la première étape du pas de temps suivant.

Les performances de ce couplage entre résolution d'un problème de diffusion et calculs d'équilibres thermodynamiques locaux en chaque nœud du maillage et à chaque instant sont pénalisés par l'appel direct au code de calcul thermodynamique. Par conséquent, pour réduire le surcoût engendré par les calculs thermodynamiques dans la simulation globale, nous développons un réseau de neurones pour se substituer au calcul des potentiels et des mobilités.

## 2.2 Développement du réseau de neurones

Pour générer les données d'entraînement, nous avons lancé 1000 simulations d'inter-diffusion couplées au code de thermodynamique OpenCalphad. Pour toutes les simulations, nous avons imposé une symétrie coaxiale ainsi qu'un profil de température parabolique. Nous avons utilisé une grille de 40 points et avons exécuté la simulation pendant 1000 pas de temps, chacun représentant 0.5 s, pour un total de 500 s. Chaque simulation débute avec un mélange uniforme d'Uranium, de Plutonium et d'Oxygène à chaque point de la grille et la température minimale du système $T_{\mathrm{min}}$, que l'on doit spécifier.

Pour chaque simulation, nous avons donc tiré alétoirement :
- La composition initiale :
  - $N_{\mathrm{U}}^0 \sim U(0.6, 0.9)$
  - $N_{\mathrm{O}}^0 \sim U(1.975, 1.999)$

— $N^0_{\mathrm{Pu}} = 1 - N^0_U$
— La température minimale du système : $T_{\min} \sim U(800, 1100)$

Les données utilisées pour l'entraînement du réseau sont les paires entrée-sortie produites par OpenCalphad au cours de la simulation. 70 % des simulations sont utilisées pour l'entraînement, 15 % pour la validation et 15 % pour les tests.

Le réseau de neurones que nous avons développé est un réseau complètement connecté comprenant 16 couches, chacune de 512 neurones utilisant une connexion de saut (*skip connection*) à chaque couche et avec une fonction d'activation Tanh. Nous avons utilisé une fonction de perte $\mathcal{L}$ incorporant la perte Mean Squared Error (MSE) classique, notée $\mathcal{L}_\alpha$, et un terme supplémentaire $\mathcal{L}_\beta$ qui prend en compte plus spécifiquement les différences de potentiels chimiques et termes d'inter-diffusion apparaissant dans le système Eqs1.

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}_\alpha(y, \hat{y}) + \mathcal{L}_\beta(y, \hat{y}),$$

Avec :

$$\mathcal{L}_\alpha(y, \hat{y}) = \sum_{e \in \{\mathrm{O,Pu,U}\}} ((\mu_e - \hat{\mu}_e)^2 + (M_e - \hat{M}_e)^2),$$

et

$$\mathcal{L}_\beta(y, \hat{y}) = \frac{1}{2} \sum_{\substack{e,e' \in \{\mathrm{O,Pu,U}\} \\ e \neq e'}} (((\mu_e - \mu_{e'}) - (\hat{\mu}_e - \hat{\mu}_{e'}))^2 + (M_{e,e'} - \hat{M}_{e,e'})^2).$$

Sur la base de test, nous avons obtenu des résultats prometteurs avec un coefficient de détermination $R^2$ supérieur à 99,99% sur toutes les quantités prédites. Nous montrons la pertinence de notre approche en déployant le modèle sur une simulation de 500 s, voir Figure 3.

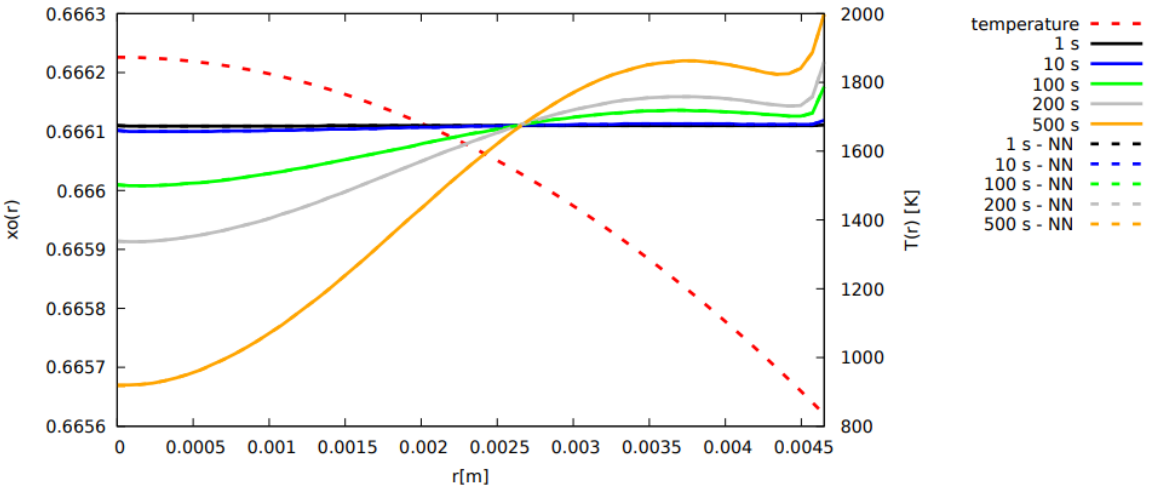



FIGURE 3 – Instantanés du profil de fraction molaire d'oxygène ($x_{\mathrm{O}}$) par rapport au rayon du granule ($r[m]$). La ligne en pointillés rouges montre le profil de température fixe. Les lignes continues (resp. en pointillés) montrent l'évolution des fractions molaires dans la simulation couplée avec OpenCalphad (resp. le modèle de réseau de neurones). Les lignes continues et en pointillés sont superposées du fait que l'utilisation du réseau de neurones fournit des résultats très proches de ceux avec appel à OpenCalphad.

# 3 Détection des données hors du domaine du modèle

## 3.1 Méthodologie de détection

Malgré les performances de notre réseau de neurones obtenues sur les données de test, il est nécessaire de fournir des garanties sur sa fiabilité dans le cas d'un déploiement en production. La détection des données hors du domaine [1] du modèle [2] est une tâche importante pour répondre à cet objectif. Elle consiste à identifier les données d'entrée qui se situent en dehors du domaine où il est capable de fournir des prédictions correctes et fiables. Cela permet de rejeter le résultat du réseau de neurones si on considère que le niveau de confiance accordé à la prédiction n'est pas suffisamement élevé.

Notre objectif est de développer un système de surveillance en temps réel [2] pour la détection des entrées hors du domaine du modèle pour plusieurs raisons. Un système de surveillance en temps réel a pour but de détecter et de répondre rapidement aux déviations ou aux anomalies pendant l'exécution de la simulation, garantissant ainsi sa pertinence et son efficacité alors que le comportement du système physique évolue.

La méthode de surveillance que nous proposons est basée sur les boîtes englobantes (*Bounding boxes*), sur une approche similaire à celle proposée dans [3]. L'idée de base est de regrouper les données d'entraînement en clusters en fonction des valeurs des neurones avant activation des couches cachées du réseau. Puis, on établit le domaine d'entrée acceptable en définissant des bornes inférieures et supérieures sur les états des neurones pour chacun de ces clusters, construisant ainsi les boîtes englobantes. Le système de surveillance, ainsi construit dans [3], rejette une entrée si l'ensemble des états des neurones des couches cachés résultant de cette entrée ne se situent dans les bornes d'aucun cluster.

Cette méthode est très conservative, au sens où un seul dépassement des boîtes englobantes pour un seul neurone est rédhibitoire. Nous proposons une extension de cette méthode, nommée **BB-AS** pour *Bounding box with Anomaly Score*, qui attribue un score d'anomalie consistant à calculer le nombre de dépassements des bornes definies sur les états des neurones sur chaque cluster et à conserver le nombre minimum de dépassement sur tous les clusters. Ainsi, le score d'anomalie d'une entrée varie entre 0 (lorsque l'entrée se situe dans les bornes d'un des clusters), et le nombre de neurones des couches cachées de réseau (lorsque l'entrée a dépassé tous les seuils de tous les clusters). La méthode BB-AS offre alors une souplesse qui permet de prendre en compte les cas où les données se situant hors de quelques bornes des boîtes englobantes peuvent malgré cela être pertinentes.

1. Cette tâche se distingue de l'approche plus classique qui consiste à identifier les données dites hors distribution. Celle-ci cherche en effet à détecter les données qui ne sont pas conformes à la distribution attendue ou normale de l'ensemble de données d'entraînement, mais cette fomulation est ambiguë, car il est souvent difficile de définir avec précision ce qui constitue une donnée anormale.

## 3.2 Exemple d'application de la méthode BB-AS au métamodèle d'OpenCalphad

Afin de tester la méthode BB-AS sur le problème présenté dans la section 2, nous avons utilisé 1000 simulations d'inter-diffusion couplées au code de thermodynamique OpenCalphad. . Pour ces simulations, nous nous sommes placés dans la même configuration que celle décrite en section 2. La principale différence est que nous avons élargi la distribution des paramètres initiaux de simulation afin d'inclure à la fois des données hors et dans le domaine du modèle. Le tirage aléatoire de ces paramètres est modifié comme suit :

— Composition initiale :
  — $N_{\mathrm{U}}^0 \sim U(0.4, 0.95)$
  — $N_{\mathrm{O}}^0 \sim U(1.95, 1.999)$
  — $N_{\mathrm{Pu}}^0 = 1 - N_{\mathrm{U}}^0$
— La température minimale du système : $T_{\min} \sim U(750, 1100)$

Dans cette démarche, nous avons repris le même modèle préalablement entraîné.

Pour définir le domaine du modèle, nous avons utilisé un seuil supérieur sur la perte $\mathcal{L}$ de $10^{-4}$, calculée à partir des paires entrées-sorties produites par OpenCalphad et des prédictions du réseau. Nous considérons qu'une donnée dont la perte dépasse ce seuil est une donnée hors du domaine de modèle. La valeur de ce seuil a été définie empiriquement en remarquant que des simulations comportant des erreurs inférieures $10^{-4}$ ne divergent pas. Ainsi, pour construire nos boîtes englobantes, nous avons utilisé les données de l'ensemble d'entraînement en rejetant les instances où la perte dépasse ce seuil. Cette expérience n'a nécessité que l'utilisation d'un seul cluster pour regrouper les données, ce qui conduit finalement à une seule boîte englobante.

Nous avons comparé la méthode BB-AS à la méthode de référence *Deep ensemble* [7], en construisant un ensemble profond composé de 8 modèles de même architecture que notre modèle et entraîné sur les mêmes données. Nous avons utilisé le logarithme décimal [2] de la moyenne de la variance des sorties des huit réseaux comme critère de détection de données hors domaine du modèle. Nous avons également comparé la méthode BB-AS à l'approche originale [3], qui ne nécessite pas de calibration.

La comparaison des performances a été effectuée vis-à-vis de la capacité à accepter ou rejeter correctement des données hors du domaine du modèle, c'est-à-dire, menant à une perte $\mathcal{L}$ supérieure à $10^{-4}$. Nous calculons pour ce faire les métriques conventionnelles FPR@95 (False Positive Rate at 95% True Positive Rate), AUROC (Area Under the Receiver Operating Characteristic curve) et AUPR (Area Under the Precision-Recall curve).

La méthpode BB-AS présente une amélioration significative par rapport à la méthode originale [3], qui rejette environ 42% des données situées dans le domaine du modèle dans notre cas de test. Les performances de la méthode BB-AS sont légèrement moins satisfaisantes que celles obtenues avec la méthode d'ensembles. Cependant, elle se distingue par le fait qu'elle ne nécessite que l'entraînement et l'inférence d'un seul modèle, le calcul du score d'anomalie ajoute une durée négligeable. La méthode d'ensemble nécessite l'entraînement puis l'inférence

2. Cette transformation était nécessaire car les valeurs de variance variaient sur des échelles différentes sur l'ensemble de test.

| Méthode | FPR@95 | AUROC | AUPR | Temps de calcul |
|---|---|---|---|---|
| Méthode de [3] | 0.420 | 0.790 | 0.866 | $\approx \mathbf{1}$ |
| Méthode BB-AS | 0.053 | 0.991 | 0.992 | $\approx \mathbf{1}$ |
| Méthode d'ensemble profond | **0.025** | **0.996** | **0.996** | 8 |

TABLE 1 – Comparaison des performances entre les trois méthodes pour la détection de données hors domaine. Le temps de caclul est donné par référence à l'exécution d'un seul modèle neuronal.

de plusieurs modèles (huit dans notre cas). Cela ajoute à la fois une complexité supplémentaire au processus d'entraînement et une augmentation significative du temps de calcul à l'inférence, qui diminue les bénéfices apportés par l'utilisation d'un modèle de substitution à un code coûteux. Cet aspect est important à considérer pour notre cas d'utilisation.

### 3.3 Perspectives

Ce succès constitue un premier pas dans une approche à complexité croissante, et nous prévoyons d'étendre notre étude à des cas de dimensionnalité plus élevée à l'avenir. En continuant à explorer et à développer la méthode BB-AS, nous visons à adresser des problèmes plus complexes et à obtenir des résultats encore plus significatifs. Cette extension à des cas de plus grande dimensionnalité nous permettra d'évaluer la robustesse et la généralisation de notre approche dans des situations plus diversifiées et réalistes. En fin de compte, notre objectif est de proposer une solution innovante et efficace pour la détection des données hors domaine du modèle, tout en minimisant les exigences en termes de ressources et de temps de calcul.

## Bibliographie


[1] Alison Cozad, Nikolaos V. Sahinidis, and David C. Miller. Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6) :2211–2227, 2014.

[2] Joris Guerin, Kevin Delmas, Raul Ferreira, and Jérémie Guiochet. Out-of-distribution detection is not all you need. *AAAI Conference on Artificial Intelligence*, pages 14829–14837, 2023.

[3] Thomas A. Henzinger, Anna Lukina, and Christian Schilling. Outside the box : Abstraction-based monitoring of neural networks. *CoRR*, abs/1911.09032, 2019.

[4] C. Introïni, J. Sercombe, and Bo Sundman. Development of a robust, accurate and efficient coupling between pleiades/alcyone 2.1 fuel performance code and the opencalphad thermo-chemical solver. *Nuclear Engineering and Design*, 369 :110818, 2020.

[5] Clément Introïni, Jérome Sercombe, Isabelle Ramière, and Romain Le Tellier. Phase-field modeling with the taf-id of incipient melting and oxygen transport in nuclear fuel during power transients. *Journal of Nuclear Materials*, 556 :153173, 2021.

[6] Kurt and Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257, 1991.

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

[8] Bo Sundman, Ursula R Kattner, Mauro Palumbo, and Suzana G Fries. OpenCalphad-a free thermodynamic software. *Integrating Materials and Manufacturing Innovation*, 4(1) :1, 2015.

[9] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Michal Walczak, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, page 1–1, 2021.

# Effet de la complexité du réseau LSTM sur l'explicabilité en Maintenance Prédictive

Lamine NDAO [1,2] & Genane YOUNESS [1,2] & Ndeye NIANG [2] & Gilbert SAPORTA [2]

[1] *Laboratoire LINEACT CESI, Nanterre, IDFC*
[2] *Laboratoire Cedric-MSDMA-CNAM, Paris, France*
*{mlndao ; gyouness}@cesi.fr ; {n-deye.niang ; gilbert.saporta}@cnam.fr*

**Résumé.** La nature complexe des données en maintenance prédictive impose souvent l'utilisation de modèles d'apprentissage profonds. Malgré leur efficacité dans la prédiction du RUL (durée de vie résiduelle des machines), ces « boites noires » fournissent des résultats qui ne sont pas directement compréhensibles. Ainsi, des méthodes XAI post hoc sont généralement utilisées pour les expliquer. La modélisation inclut habituellement le choix de la complexité du modèle telle que la profondeur du réseau. Par ailleurs, on pourrait se demander si une complexité élevée du modèle ne freine pas la capacité des méthodes XAI dans l'explication des prédictions. Cette étude examine l'effet de la profondeur du réseau LSTM sur la qualité des explications des méthodes XAI, post hoc, locales, LIME, SHAP et L2X, utilisant huit métriques d'évaluation. Les résultats obtenus montrent que la qualité des explications peut suivre une certaine tendance en fonction de la complexité du réseau et selon la propriété de l'explication évaluée. Ces résultats ont montré également le manque de concordance entre les métriques d'évaluation, impliquant ainsi un besoin de cadre consensuel plus fiable dans l'évaluation des méthodes XAI.

**Mots-clés** Maintenance prédictive, Méthodes post-hoc locales, XAI, LSTM, métriques d'évaluation de XAI.

**Abstract.** The complexity of predictive maintenance data often requires the use of deep learning models. Despite their effectiveness in predicting RUL (Remain Useful Life), these "black boxes" provide results that are not directly comprehensible. Thus, post-hoc XAI methods are generally used to explain them. Modeling usually includes the choice of model complexity, such as network depth (number of layers). On the other hand, it may be asked whether high model complexity inhibits the ability of XAI methods to explain predictions. This study examines the effect of LSTM network depth on the explanation quality of XAI, post hoc, local, LIME, SHAP and L2X methods, using eight evaluation metrics. The results obtained show that the quality of explanations can follow a certain trend depending on the complexity of the network and the property of the explanation being evaluated. These results also showed a lack of agreement between evaluation metrics, implying a need for a more reliable consensus framework in the evaluation of XAI methods.

**Keywords.** Predictive maintenance, local post-hoc methods, XAI, LSTM, XAI evaluation metrics.

# Introduction et travaux antérieurs

En maintenance prédictive, les réseaux de neurones récurrents de type LSTM sont de plus en plus utilisés. Cela peut s'expliquer par leurs bonnes performances dans ce domaine (Gouriveau et al. (2013)). Cependant, malgré ces performances grandissantes, ces méthodes sont souvent considérées comme des "boîtes noires" en raison de leurs structures internes complexes. Cette dernière implique le manque de transparence dans le processus de prédiction qui ne permet pas une explication directe des prédictions comme dans le cas d'une régression linéaire. D'ailleurs, ce manque de transparence a suscité de nombreuses questions relatives à la confiance en l'IA. En réponse, l'IA eXplicable (XAI) a été présentée comme une solution (DARPA Gunning and Aha (2019)) Depuis, une diversité de méthodes pour expliquer les résultats de ces modèles qualifiés de "boites noires" a été proposée. Ces méthodes peuvent être classées en différentes catégories, selon leur approche, le type d'explication qu'elles fournissent ou la portée de l'explication fournie. Ainsi, elles peuvent être : intrinsèques ou post hoc, selon qu'elles interviennent pendant ou après l'apprentissage du modèle ; globales, de cohorte ou locales, selon leur capacité à expliquer la prédiction d'une observation, d'un groupe d'observations ou l'ensemble des observations. Plusieurs auteurs soutiennent la pertinence des méthodes XAI. Toutefois, on peut s'interroger sur la fiabilité de ces méthodes XAI. On peut également se demander comment évaluer qualitativement et quantitativement leurs résultats et sur quelle base elles pourront être comparées. Pour répondre à ce besoin, diverses approches ont été proposées. Ces approches comprennent des méthodes qualitatives basées sur l'appréciation humaine ainsi que des approches quantitatives visant à évaluer quantitativement certaines des propriétés qu'une explication devrait satisfaire. Cette évaluation quantitative se fonde sur l'analyse de la relation entre les données et les explications (Honegger (2018)) ou entre les données, les prédictions et les explications (Solís-Martín et al. (2023)). Lorsqu'on parcourt la littérature, on constate que peu d'études s'intéressent à la relation entre la qualité des explications fournies par une méthode XAI et la complexité du modèle d'analyse (ex. nombre de couches, nombre de cellules de neurones). Dans le cadre des réseaux de neurones, on sait qu'augmenter le nombre de couches cachées (rajouter de la complexité) peut améliorer la performance du modèle d'analyse. Ainsi, on peut se demander si ce rajout de complexité ne limite-t-il pas la capacité d'une méthode XAI dans le processus d'extraction d'explications, sachant que les méthodes XAI fournissent un lien approximatif entre les prédictions du modèle d'analyse et les variables.

Dans cette étude, on se propose d'analyser empiriquement le lien entre la complexité d'un modèle d'analyse de type LSTM et la qualité des explications fournies par trois méthodes XAI LIME (Ribeiro et al. (2016)), SHAP (Lundberg and Lee (2017)) et L2X (Chen et al. (2018)) évaluée au moyen de huit métriques d'évaluation.

# 1 Méthodologie

**Notations :**

- N : Nombre d'observations (nombre de moteurs)
- $X = (x_i^t)_{(i \in N, t \in T)}$ l'ensemble des observations avec $i$ le moteur et $t$ une date donnée.

— $Y_t$ : le RUL observée à la date $t$
— $f$ : la fonction de prédiction du modèle d'analyse
— $\epsilon = \{\epsilon_i\}_{(i \in N)}$ : poids des variables ("feature importance) dans l'explication de la prédiction du $\mathbf{RUL}_i$.
— $\rho$ : coefficient de Spearman

**Long Short-Term Memory (LSTM) :** LSTM (Hochreiter and Schmidhuber (1997)) est un réseau de neurones de type récurrent. Il s'agit d'un modèle performant dans le traitement des données temporelles avec des structures complexes, résolvant le problème du gradient optimal de la rétro propagation pour modifier les poids du réseau.

**L'eXplicabilité en Intelligence Artificielle (XAI) :** Dans le contexte de la prédiction du RUL, on s'intéresse particulièrement aux parties du moteur qui sont responsables de la dégradation de la durée de vie résiduelle d'un moteur donné. Ainsi, dans notre analyse, nous nous concentrerons sur trois méthodes XAI locales post hoc : LIME (Local Interpretable Model-Agnostic Explanations), SHAP et L2X (Learning to Explain) qui sont considérées comme des méthodes d'explication locale basées sur les perturbations. Nous notons $(x, y)$ une observation dans $(X, Y)$, et $g$ la fonction d'apprentissage du modèle de substitution (e.g. la régression linéaire).

— Pour l'approche LIME, l'idée principale est de créer un ensemble d'observations à partir de $X_i$ à partir de la distribution de $h$. Ensuite, un modèle linéaire $g$ est entrainé sur cet échantillon avec une contrainte d'éparpillement. Enfin, les coefficients de régression $phi_i$ sont utilisés comme l'effet des différentes variables impliquées dans la prédiction.
— KernelSHAP utilise la valeur Shapley issue de la théorie des jeux pour attribuer une valeur, appelée valeur SHAP, à chaque variable, décrivant sa contribution à la prédiction finale. Par souci de simplicité, nous écrirons SHAP en référence à KernelSHAP.
— L2X cherche le sous-ensemble de variables le plus informatif en termes de prédiction correspondante pour cette instance. Le sous-ensemble est déterminé par un sélecteur de variables, par approximation variationnelle, qui est optimisée de manière à maximiser l'information mutuelle $MI$ entre les caractéristiques et la prédiction correspondante.

En général, la génération des explications est basée sur la perturbation des variables et la sélection de voisinage. Dans le contexte de cette étude, étant donné que nous traitons de séries temporelles, nous adoptons l'approche de perturbation proposée par Solís-Martín et al. (2023), qui est plus appropriée pour les séries temporelles. La qualité des explications générées est évaluée par des métriques d'évaluation XAI qui vérifient certaines propriétés que ces explications doivent respecter, telles que la robustesse ou la stabilité.

**Évaluation des méthodes XAI :** Doshi-Velez and Kim (2017) ont décrit trois catégories d'approches d'évaluation pour des méthodes XAI : **"Human-grounded Evaluation"** qui englobe les méthodes basées sur l'appréciation humaine ; **"Application-grounded Evaluation"** qui implique des approches basées sur l'appréciation humaine spécifique à une application particulière, avec un accent prédominant sur les opinions d'experts dans le domaine concerné et **"Functionally-grounded Evaluation"** qui concerne les approches utilisant

des fonctions mathématiques ou proxy pour évaluer quantitativement la qualité des modèles post-hoc.

Dans ce travail, on s'intéresse particulièrement à la dernière approche d'évaluation. Il s'agit de métriques ayant comme but d'évaluer certaines propriétés d'une "bonne explication". Par exemple, on pourrait s'intéresser à la robustesse d'une explication. Autrement dit, on pourrait voir si les observations qui se ressemblent ont tendance à avoir des explications de leur prédication ressemblantes. La Table 1, présente les huit métriques utilisées dans cette étude. Un travail d'état de l'art nous a permis de réaliser ce tableau. Pour chaque métrique, nous avons fourni une description détaillée, sa base théorique et les propriétés qu'elle cherche dans l'évaluation de la performance des méthodes XAI.

| Métrique | Propriétés | Formule | Description |
|---|---|---|---|
| **Identité** | Fidélité | $d(x_i, x_j) = 0 \Longrightarrow d(\epsilon_i, \epsilon_j) = 0$ | Deux observations identiques au regard de $d$ doivent recevoir des explications identiques au regard de $d$ aussi (Honegger (2018)). |
| **Séparabilité** | Fidélité | $(x_i, x_j) \neq 0 \Longrightarrow d(\epsilon_i, \epsilon_j) > 0$ | Deux observations $(i, j)$ différentes au regard de la distance $d$ ne peuvent pas recevoir 2 explications identiques au regard de $d$(Honegger (2018)). |
| **Stabilité** | Précision/Fidélité | $\rho_i = \rho(XX_i, E_i)$, $\rho_i > 0 \ \forall i \in N$ | La stabilité évalue si 2 observations similaires en termes de variable explicative $X$ sont également similaires en termes d'explications $\epsilon$. Un $\rho_i$ élevé indique une interprétation plus intuitive(Honegger (2018)). |
| **Congruence** | Cohérence | $\delta = \sqrt{\frac{\sum(\alpha_i - \bar{\alpha})^2}{N}}$ | La congruence évalue la variabilité de la cohérence. Une valeur $\delta$ plus petite indique une cohérence plus stable dans les prédictions (Doshi-Velez and Kim (2017)). |
| **completeness** | Représentativité | $\gamma_i = \frac{e_e^i}{p_e^i}$ | La complétude évalue le ratio entre l'erreur de prédiction initiale et l'erreur après une perturbation des données initiales. Une valeur proche de 1 indique une meilleure qualité de l'explication (Doshi-Velez and Kim (2017)). |
| **Cohérence** | Cohérence | $\alpha_i = \lvert p_e^i - e_e^i \rvert$ | La cohérence évalue la différence entre l'erreur de prédiction avec les données réelles et l'erreur de prédiction après perturbation des variables non importantes selon la méthode XAI. Une faible valeur d'$\alpha$ indique une cohérence élevée dans l'approche d'explicabilité (Doshi-Velez and Kim (2017)). |
| **Acumen** | Robustesse | $\omega = 1 - \frac{\sum_{f_i \in \mathcal{I}} \frac{p_a(f_i)}{N}}{M}$ | L'acumen évalue si l'importance d'une variable selon la méthode XAI dépend de sa position dans les données. Une valeur d'acumen élevée indique une stabilité de l'importance des variables(Solís-Martín et al. (2023)). |
| **Sélectivité** | Sélectivité | | La selectivité évalue la perturbation des variables les plus importantes. Elle consiste à ordonner les variables selon leur importance selon une méthode d'explicabilité, puis à perturber les données en substituant des variables aléatoires pour les variables les plus importantes, et enfin à calculer l'erreur de prédiction pour chaque perturbation (Laugel (2020)). |

TABLE 1 – L'ensemble des métriques d'évaluation utilisées dans cette analyse

# 2 Expérimentations

**Données :** Cette analyse s'est basée sur les données C-MAPSS (Commercial Modular Aero-Propulsion System Simulation (Saxena et al. (2008)). Ces données enregistrent la durée de fonctionnement jusqu'à la défaillance des moteurs d'avions en simulant un large éventail de conditions opérationnelles réalistes, de paramètres de défaillance et de tendances de

dégradation dans différentes sections du système moteur.

La Figure 1 montre une vue abstraite de haut niveau de l'ensemble de données C-MAPSS. Essentiellement, les données de chaque moteur sont des séries temporelles multivariées (STM) qui consistent en des mesures prises au fil du temps à partir de différents capteurs montés sur le moteur. Chaque intervalle de temps correspond à un cycle de fonctionnement du moteur. L'objectif est de prédire la durée de vie utile restante (RUL), c'est-à-dire le nombre de cycles de fonctionnement restants pour le moteur, compte tenu de l'historique des mesures de ses capteurs. Les données de chaque moteur sont représentées à l'aide d'un total de 21 capteurs et de 3 modes de fonctionnement. Elles comprennent 4 parcs (flottes) de moteurs FD01, FD002, FD003, FD004. On considérera le groupe de moteurs FD004 dans cette analyse. Alors que les données d'entrainement enregistrent les parcours jusqu'à la défaillance, les données de test contiennent les mesures historiques des capteurs des moteurs jusqu'à un certain moment, avec une durée de vie résiduelle connue.

**Pré-traitement :** Dans cette étude, le prétraitement des séries temporelles est effectué en trois phases : lissage exponentiel, fenêtre temporelle et RUL rectifiée.

**Normalisation :** On commence par une normalisation des données en utilisant l'approche **min-max** dans chaque groupe de moteurs ayant les mêmes conditions opérationnelles. Cela permet de mettre les données issues des différents capteurs sur une même échelle. La formule est donnée par :

$$\text{norm}\left(x_{i,j}^{t}\right) = \frac{x_{i,j}^{t} - \min\left(x_j\right)}{\max\left(x_j\right) - \min\left(x_j\right)} - 1 \tag{1}$$

**Lissage exponentiel des données :** Pour produire une estimation précise de la RUL malgré la présence de bruit dans les données, un processus de lissage exponentiel est appliqué. Le lissage exponentiel attribue différents poids aux observations historiques en fonction de leur récence. Le choix du paramètre de lissage $\alpha$ dans le lissage exponentiel détermine le niveau d'importance accordé aux observations récentes. Les valeurs ajustées utilisent le paramètre de lissage $\alpha$ selon l'équation suivante :

$$\hat{y}_{t+1|t} = \alpha y_t + (1-\alpha)\hat{y}_{t|t-1} \tag{2}$$

$\alpha = 1$ indique un apprentissage rapide, ce qui signifie que les prévisions sont basées sur les valeurs les plus récentes, tandis que $\alpha = 0$ indique un apprentissage lent.

**Fenêtre temporelle et RUL rectifié :** Après avoir réduit le bruit par lissage exponentiel, une fenêtre temporelle glissante de longueur fixe $TW$ est appliquée pour convertir les données de séries temporelles multivariées. En fixant une durée de vie résiduelle à un certain seuil dit RUL rectifié $RUL_{early}$, le système est considéré comme "sain" jusqu'à ce qu'il atteigne ce point prédéfini. Cela permet au modèle de se concentrer sur l'apprentissage à partir du cycle $RUL_{early}$, quelle que soit la durée de vie antérieure du moteur. Cette valeur est à définir de façon optimale par le concepteur.

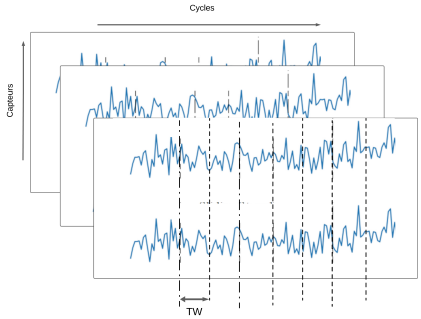



FIGURE 1 – Format des données et leur découpage en séquence suivant $TW$ fixé

**Processus d'analyse :** L'objectif de cette étude est d'analyser empiriquement l'effet de la complexité du modèle d'analyse LSTM sur la qualité des explications fournies par les méthodes LIME, KernelSHAP et L2X. Pour ce faire, nous avons adopté le processus suivant :

1. Un modèle LSTM à une couche a été optimisé. Cette optimisation a été faite en choisissant les paramètres de prétraitement ($\alpha$, $TW$ et $RUL_{early}$) qui donnaient les meilleures performances en termes des deux critères $S-score$ et $RMSE$.
2. On répète 5 fois les étapes suivantes :
   — Une explication des prédictions sur un échantillon de 20 moteurs est obtenue à l'aide des 3 méthodes LIME, KernelSHAP et L2X.
   — L'évaluation des explications fournies est faite en utilisant les 8 métriques présentées dans la Table 1
3. L'étape (2) est répétée en ajoutant à chaque fois une couche aux réseaux de neurones LSTM jusqu'à avoir un modèle à 4 couches.

# 3 Résultats et discussions

On commencera par présenter les performances des modèles d'analyse. Ensuite, on présentera la qualité des explications en fonction de la profondeur du réseau LSTM. Le dernier paragraphe portera sur l'analyse du lien entre les métriques.

**Performances versus profondeur du réseau LSTM :** Une première étape de cette analyse portait sur l'optimisation du modèle à 1 couche en choisissant les paramètres de prétraitement optimaux. Ainsi, nous avons retenu le triplet de paramètres ($TW = 40$, $\alpha = 0.5$, $RUL_{early} = 100$) qui a donné les meilleures performances en termes de $RMSE$ et $S-score$. La structure du modèle d'analyse de base est présentée dans la Table 2. La Table 3 montre la performance du modèle en fonction du nombre de couches cachées. L'analyse de cette Table montre qu'en passant de 1 à 2 couches cachées, on arrive à obtenir de meilleures performances : $RMSE$ qui passe de 13.58 à 9.88 ; $S-score$ passe de 832.8 à 445. Cependant, cette amélioration du modèle n'est pas observée lorsqu'on passe de 2 à 3 couches cachées ou de 3 à 4 couches cachées. Ceci permet de noter que, dans le cadre de cette étude, le meilleur

modèle en termes de performances, n'est pas le modèle ayant le plus grand nombre de couches cachées.

| **Hyperparamètres** | **valeurs** |
|---|---|
| Couches | 1 |
| Nœuds | 64 |
| Dropout | 0.2 |
| Batch-size | 120 |
| Learning-rate | $10^{-3}$ |

TABLE 2 – Les hyper-paramètres de LSTM

| Couches | RMSE | S-score |
|---|---|---|
| 1 | 13.58 | 832.8 |
| 2 | **9.88** | **445.61** |
| 3 | 10.56 | 479.67 |
| 4 | 10.22 | 515.54 |

TABLE 3 – Performance des modèles en fonction du nombre de couches cachées

Cependant, plus de couches cachées, implique une liaison moins linéaire entre la variable cible $y$ et les variables dépendantes $X$. De surcroît, certaines méthodes XAI comme LIME, lient, de façon linéaire, $X$ et $y$. LIME se base sur un modèle de substitution comme la régression linéaire à la place du modèle d'analyse afin d'utiliser les coefficients de régression pour expliquer l'influence de chaque $X_j$ dans la prédication de $\bar{y}$. Ainsi, on se questionne sur la fiabilité des explications fournies par ces méthodes dites post hoc. Pour avoir une réponse de façon empirique, nous avons évalué la qualité des explications fournies par 4 modèles avec des nombres de couches cachées différentes (1 à 4) sur la base de 8 métriques d'évaluation.

**Complexité du réseau versus qualité des explications :** Les résultats de cette évaluation (Table 4) montre que, de manière générale, SHAP fournit de meilleurs résultats au regard des métriques d'évaluation utilisées. En effet, elle donne une meilleure qualité des explications au regard de 5 des 8 métriques d'évaluation utilisées (Cohérence : 0.20, Congruence : 0.24, Sélectivité : 0.79, Acumen : 0.5 ; Congruence 0,42). On note également, qu'en général, la valeur de la métrique sélectivité augmente lorsqu'on augmente le nombre de couches cachées. On peut le voir sur les méthodes comme LIME (1 couche : 0.58, 2 couches : 0.72, 3 couches : 0.73, 4 couches : 0.78). Cette tendance est observée sur l'ensemble des 3 méthodes XAI.

Nous avons également analysé le comportement des valeurs prises par les 6 métriques d'évaluation en excluant Id et Sep, vu qu'elles ne présentent pas variations (Figure 2). Dans cette analyse, on cherche à déceler empiriquement l'effet de la profondeur du modèle (complexité du modèle) sur la qualité des explications fournies. L'analyse montre que pour certaines métriques (e.g. sélectivité), on arrive à déceler une tendance de la qualité des explications dans certaines méthodes XAI lorsque le nombre de couches augmente. Cela montre que la performance du modèle d'analyse n'est forcément liée à la qualité de l'explication. De même, pour certaines méthodes XAI, (ex. SHAP), lorsqu'on augmente le nombre de couches, la qualité de l'explication s'améliore selon certaines métriques (ex. Sélectivité), et se dégrade selon d'autres métriques (cohérence, completeness, congruence).

| Méthodes | Couches | Id ↑ | Sep ↑ | St ↑ | Co ↓ | Con ↓ | Sel ↑ | Ac ↑ | Com ↑ |
|---|---|---|---|---|---|---|---|---|---|
| LIME | 1 | 1.0(0.0) | 1.0(0.0) | 0.98(0.04) | 0.24(0.11) | 0.27(0.06) | 0.58(0.04) | 0.07(0.04) | 0.14(0.2) |
| | 2 | 1.0(0.0) | 1.0(0.0) | **1.0(0.0)** | **0.22(0.08)** | **0.26(0.06)** | 0.72(0.06) | **0.12(0.01)** | 0.07(0.04) |
| | 3 | 1.0(0.0) | 1.0(0.0) | 0.98(0.04) | **0.22(0.08)** | **0.26(0.06)** | 0.73(0.06) | 0.09(0.04) | **0.21(0.38)** |
| | 4 | 1.0(0.0) | 1.0(0.0) | **1.0(0.0)** | 0.22(0.09) | 0.26(0.06) | **0.78(0.06)** | 0.09(0.03) | 0.19(0.35) |
| SHAP | 1 | 1.0(0.0) | 1.0(0.0) | **0.96(0.05)** | **0.20(0.08)** | **0.24(0.06)** | 0.64(0.06) | 0.42(0.12) | **0.42(0.1)** |
| | 2 | 1.0(0.0) | 1.0(0.0) | 0.88(0.16) | 0.22(0.08) | 0.26(0.06) | 0.77(0.07) | **0.50(0.06)** | 0.19(0.13) |
| | 3 | 1.0(0.0) | 1.0(0.0) | 0.94(0.09) | 0.22(0.08) | 0.26(0.06) | 0.78(0.05) | 0.43(0.07) | 0.11(0.18) |
| | 4 | 1.0(0.0) | 1.0(0.0) | 0.94(0.09) | 0.22(0.09) | 0.26(0.06) | **0.79(0.07)** | 0.34(0.05) | 0.17(0.19) |
| L2X | 1 | 1.0(0.0) | 1.0(0.0) | 0.96(0.05) | 0.25(0.13) | **0.27(0.07)** | 0.59(0.05) | **0.04(0.02)** | **0.38(0.22)** |
| | 2 | 1.0(0.0) | 1.0(0.0) | 0.96(0.05) | **0.24(0.1)** | 0.28(0.08) | 0.72(0.06) | 0.03(0.01) | 0.21(0.36) |
| | 3 | 1.0(0.0) | 1.0(0.0) | **1.0(0.0)** | 0.26(0.09) | 0.29(0.07) | 0.68(0.06) | 0.03(0.01) | 0.06(0.07) |
| | 4 | 1.0(0.0) | 1.0(0.0) | 0.98(0.04) | 0.25(0.1) | 0.29(0.07) | **0.74(0.05)** | 0.03(0.01) | 0.27(0.41) |

TABLE 4 – Moyenne (écart-type) des métriques d'évaluation en fonction de la profondeur du réseau LSTM. Les moyennes et les écarts-type ont été obtenus en répétant 5 fois la prédiction, l'explication puis l'évaluation ( Id : Identité, Sep : Séparabilité, St : Stabilité, Co : Cohérence : Con : Congruence, Sel : Sélectivité : Ac : Acumen, Com : Completeness)

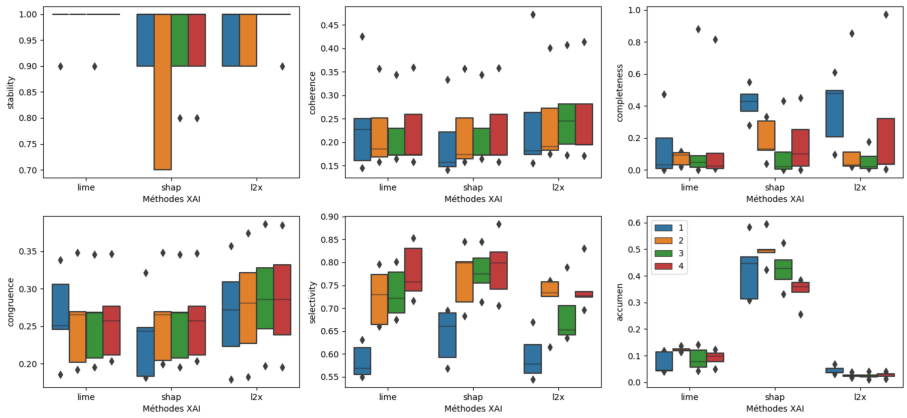



FIGURE 2 – Variation des métriques en fonction du nombre de couches cachées (couleurs), par méthode(en abscisse).

Cependant, si certaines ont décelé une tendance positive en fonction du nombre de couches cachées (Sélectivité), d'autres révèlent une tendance plutôt négative de la qualité de l'explicabilité en fonction du nombre de couches. Ceci montre qu'une complexité grandissante du modèle ne limite pas forcément la qualité de l'explication lorsque certaines propriétés de l'explication sont évaluées. Sélectivité évalue la capacité de la méthode XAI à sélectionner des variables pertinentes dans le processus de l'explication. Ainsi, on pourrait conclure qu'une complexité grandissante du modèle permet à LIME et SHAP de bien sélectionner les variables pertinentes dans le processus de l'explication.

**Relation entre les métriques d'évaluation :** L'analyse de la relation entre les métriques a permis de noter que toutes les métriques ne sont pas corrélées à un même axe (Figure 3). Ceci est compréhensible dans la mesure où elles ne sont pas censées évaluer les mêmes propriétés (Table 1). Par exemple, sélectivité évalue la capacité d'une méthode XAI à sélectionner les variables pertinentes dans l'explication des résultats d'une "boite noire",

alors que la cohérence évalue sa capacité à commettre des erreurs similaires lorsque les variables pertinentes ne sont pas perturbées (cohérence). Ainsi, on note que la cohérence et la congruence sont positivement corrélées au premier axe, car évaluant la même propriété, tandis qu'acumen, la sélectivité et la stabilité sont plutôt fortement liées au deuxième axe. La corrélation forte et positive entre acumen et sélectivité peut permettre de dire que les deux propriétés qu'elles évaluent (la robustesse et la sélectivité) peuvent être placées sur une même dimension.

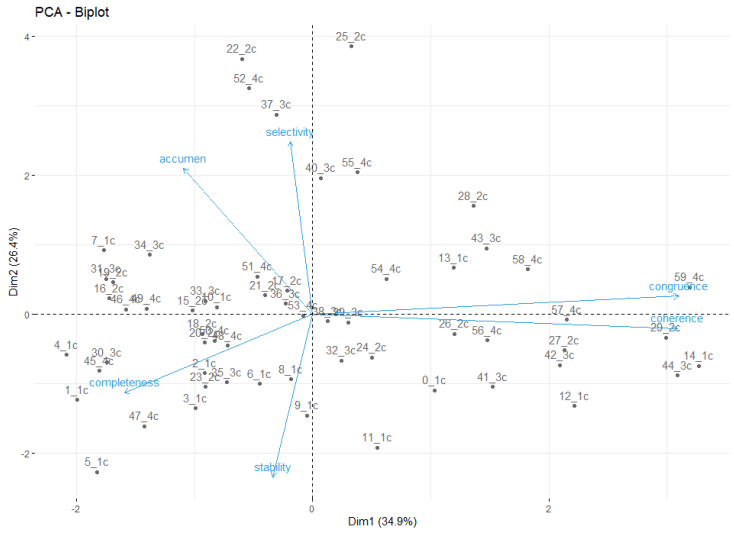



FIGURE 3 – Nuage des individus associé à l'affichage de la corrélation entre les métriques d'évaluation

# Conclusion et perspectives

Dans cette étude, il était question d'analyser l'effet de la complexité du réseau LSTM sur la qualité des explications fournies par trois méthodes XAI post-hoc LIME, SHAP et L2X, mesurée par huit proxys. Les résultats ont montré que la complexité du réseau, en termes de nombres de couches, ne limite pas en général la capacité des méthodes XAI à fournir de bonnes explications. Ainsi, il n'y a pas de compromis à faire entre la complexité du réseau en termes de nombre de couches et la transparence lorsqu'une méthode post-hoc est utilisée. Ces résultats ont également permis de noter que certaines métriques, par leur définition comme identité et séparabilité, donnaient en général une valeur de 1, et que d'autres comme la sélectivité donnait des valeurs très liées à la complexité du réseau. Par ailleurs, nos résultats ont montré une concordance entre les métriques et lient fortement et positivement deux métriques (cohérence et congruence) censées évaluer la même propriété (cohérence). Ils ont permis par contre de noter qu'une explication ne peut pas avoir à la fois les propriétés fidélité, précision et robustesse, car les métriques qui les évaluent sont directement opposées par le deuxième axe factoriel. Ceci pourrait être problématique quant au choix de la métrique à considérer pour l'évaluation des méthodes post-hoc.

Dans nos travaux futurs, on pourrait envisager de remédier à ce manquement en redéfinissant certaines métriques pour améliorer leur pertinence dans l'évaluation des méthodes XAI. Il sera également pertinent de proposer une métrique synthétique pouvant prendre en compte l'ensemble des métriques utilisées aujourd'hui. Cela permettra d'éviter le dilemme sur le choix de la métrique à considérer dans l'évaluation des méthodes XAI post-hoc en maintenance prédictive.

# Références


Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain : An information-theoretic perspective on model interpretation. In *International conference on machine learning*, pages 883–892. PMLR, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv :1702.08608*, 2017.

Rafael Gouriveau, Kamal Medjaher, Emmanuel Ramasso, and Noureddine Zehrouni. Phm – prognostics and health management de la surveillance au pronostic de défaillances de systèmes complexes. *Techniques de l'ingénieur Maintenance*, base documentaire : TIP095WEB.(ref. article : mt9570), 2013. doi : 10.51257/a-v1-mt9570. fre.

David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2) :44–58, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8) :1735–1780, 1997.

Milo Honegger. Shedding light on black box machine learning algorithms : Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv :1808.05054*, 2018.

Thibault Laugel. *Interprétabilité locale post-hoc des modèles de classification" boites noires"*. PhD thesis, Sorbonne université, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you ?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management*, pages 1–9. IEEE, 2008.

David Solís-Martín, Juan Galán-Páez, and Joaquín Borrego-Díaz. On the soundness of xai in prognostics and health management (phm). *Information*, 14(5) :256, 2023.

# Statistique bayésienne

# Prior de référence sous contraintes selon différentes mesures de dissimilarité

Antoine Van Biesbroeck [1,2], Clément Gauchy [3], Cyril Feau [2], Josselin Garnier [1]

[1] *CMAP, CNRS, École polytechnique, Institut polytechnique de Paris, 91120 Palaiseau, France ; email :* antoine.van-biesbroeck@polytechnique.edu.
[2] *Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191 Gif-sur-Yvette, France.*
[3] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191 Gif-sur-Yvette, France.*

**Résumé.** La théorie des priors de référence propose une solution à la question du choix du prior en analyse bayésienne, en construisant ce dernier comme celui qui minimise son influence a posteriori. Bien que permettant la construction d'un cadre alors qualifié d'objectif, l'expression du prior de référence prend souvent une forme proche d'un prior non informatif de Jeffreys. Cette dernière est parfois perçue comme encombrante à l'implémentation et rigide quant à l'introduction d'un jugement a priori. Dans cette communication, nous nous appuyons sur des travaux récents de la littérature, qui étendent la théorie des priors de référence. Leur formalisme et leurs résultats nous permettent en effet de définir un prior à la fois dit de référence et qui satisfait un certains nombres de contraintes pour lesquels nous définissont un cadre adéquat. Le résultat théorique que nous démontrons propose l'expression de ce prior de référence sous des contraintes linéaires, dont le choix pratique reste large et inclusif.

**Mots-clés.** Prior de référence, information mutuelle, $f$-divergence, contraintes.

**Abstract.** The reference prior theory porposes a solution to the prior choice issue in Bayesian analysis. It suggests the construction of that latter as the one that minimizes its *a posteriori* influence. While it allows the construction of a framework that one could call objective, the reference prior expression often takes a form close to a non-informative Jeffreys prior. That latter is sometimes perceived as a cumberstone to implement and to be rigid with any possible introduction of prior judgements. In this paper, we take as support an extention of the reference prior theory unveiled in recent works from the litterature. Their formalism and their results permit us indeed to define a prior which both (i) is called a reference prior and (ii) satisfies certain constraints for which we difine an appropriate framework. The theoretical result that we prove expresses that reference prior under linear constraints, whose choice stays wide and inclusive.

**Keywords.** Reference prior, mutual information, $f$-divergence, constraints.

# 1 Introduction

La théorie des priors de référence s'inscrit dans le domaine de l'analyse bayésienne par la définition et l'usage d'outils de la théorie de l'information pour répondre à la question du choix de la distribution a priori.

La définition originelle des priors de référence est due à Bernardo (1979), qui, en proposant la maximisation de l'information mutuelle comme critère de choix, a orienté la question vers la recherche du prior qui minimise son influence sur la distribution a posteriori qu'il induit. Le travail de Clarke & Barron (1994) démontre rigoureusement que sous cette définition classique, le prior de référence est celui de Jeffreys, déjà plébiscité pour sa propriété d'invariance par re-paramétrisation du modèle étudié. L'emploi et la considération de cette théorie reste au cœur de la littérature. Son étude constitue une problématique récente, que ce soit son expression dans différents contextes (Muré, 2021; Keefe et al., 2019), son approximation (Gu et al., 2018) ou son implémentation (Gauchy et al., 2023; Nalisnick & Smyth, 2017).

De récents travaux ont exploré de possibles extensions de la définition des priors de référence (Van Biesbroeck, 2023). Ceux-ci, qui s'appuient sur le point de vue de l'analyse de sensibilité globale, proposent une définition plus générale du prior de référence, qui est celui qui, en un sens, maximise une mesure de l'impact qu'induit la connaissance du paramètre d'intérêt sur la distribution des observations du système étudié. Leurs conclusions appuient la robustesse de la prise en compte du prior de Jeffreys en inférence bayésienne.

Néanmoins, dans des usages pratiques, le canevas bayésien et la question ouverte du choix de la distribution a priori est parfois plébiscitée pour sa permission d'introduction d'information, que ce soit à des fins d'optimisation de ses capacités d'estimation, ou de prise en compte de jugements d'experts.

Cette communication répond à cette problématique en proposant un résultat permissif à l'introduction de contraintes variées sur une loi a priori, tout en conservant son caractère de prior de référence. En prenant pour appui le cadre étendu des priors de référence de (Van Biesbroeck, 2023) et ses résultats, nous démontrons un théorème qui donne l'expression que doit avoir, au sens de la théorie des priors de référence, un prior qui satisfait les contraintes désirées. Notre résultat, qui se limite à des contraintes prenant la forme de projections linéaires saurait s'appliquer sur l'implémentation de contraintes sur les moments a priori ou bien sur la distribution marginale, suivant des idées proposées dans (Bousquet, 2023) relatives à l'ellicitation prédictive par exemple.

Le prochaine section propose une explicitation du canevas bayésien classique que nous considérons dans ce travail. Ensuite, après un rappel du contexte de la théorie des priors de référence généralisée telle que nous la considérons, nous dévouons la section 3 à la présentation de notre résultat principal et de sa démonstration. Finalement, nous concluons notre travail en section 4.

## 2 Canevas bayésien

Le cadre bayésien considéré dans ce travail est classiquement construit. Nous en proposons une rapide revue dans cette section.

Soit $k > 1$, considérons un espace probabilisé $(\Omega, \mathscr{P}, \mathbb{P})$ sur lequel sont définis la variable aléatoire $T$, à valeur dans l'espace mesurable $(\Theta, \mathscr{T})$, et le vecteur aléatoire $\mathbf{Y} = (Y_i)_{i=1}^k$, à valeur dans l'espace mesurable $(\mathcal{Y}^k, \mathscr{Y}^{\otimes k})$.

La variable aléatoire $\mathbf{Y}$ représente les observations itératives d'un système étudié, elles sont considérées comme indépendantes et identiquement distribuées conditionnellement à $T$ :

$$\text{pour tout } B_1, \dots, B_k \in \mathscr{Y},\ \mathbb{E}[\prod_{i=1}^{k} \mathbb{1}_{Y_i \in B_i} | T] = \prod_{i=1}^{k} \mathbb{E}[\mathbb{1}_{Y_1 \in B_i} | T] \text{ p.s.} \tag{1}$$

Généralement et dans ce travail, $\Theta$ est un sous-ensemble de $\mathbb{R}^d$, $d \geq 1$, et la distribution $\pi$ de $T$ est appelée le prior. Les distributions conditionnelles $\mathbb{P}_{Y_i|T}$ pour tout $Y_i$ existent comme suit :

$$\forall \theta \in \Theta,\ \mathbb{P}_{Y_i|T=\theta} \text{ est une probabilité sur } (\mathcal{Y}, \mathscr{Y}), \tag{2}$$

$$\forall B \in \mathscr{Y},\ \mathbb{P}_{Y_i|T=\cdot}(B) \text{ est mesurable sur } \Theta, \tag{3}$$

$$\forall A \in \mathscr{T},\ B \in \mathscr{Y},\ \mathbb{P}(T \in A, Y_i \in B) = \int_A \mathbb{P}_{Y_i|T=\theta}(B) d\pi(\theta), \tag{4}$$

avec $\mathbb{P}_{\mathbf{Y}|T} = \mathbb{P}_{Y_1|T}^{\otimes k} = \mathbb{P}_{Y|T}^{\otimes k}$ résultant de (1).

On suppose également que le problème admette une vraisemblance : il existe des densités $(\ell(\cdot|\theta))_{\theta \in \Theta}$ par rapport à une mesure commune $\mu$ sur $\mathcal{Y}$ telles que

$$\forall B \in \mathscr{Y},\ \mathbb{P}_{Y|T=\theta}(A) = \int_B \ell(y|\theta) d\mu(y) \text{ pour } \pi\text{-presque tout } \theta. \tag{5}$$

Ceci permet la définition des densités marginales et a posteriori, respectivement définies par rapport à $\mu^{\otimes k}$ et à $\pi$ :

$$\forall \mathbf{y} \in \mathcal{Y}^k,\ p_{\mathbf{Y}}(\mathbf{y}) = \int_\Theta \prod_{i=1}^{k} \ell(y_i|\theta) d\pi(\theta) = \int_\Theta \ell_k(\mathbf{y}|\theta) d\pi(\theta), \tag{6}$$

$$\forall \theta \in \Theta,\ \mathbf{y} \in \mathcal{Y}^k,\ p(\theta|\mathbf{y}) = \frac{\ell_k(\mathbf{y}|\theta)}{p_{\mathbf{Y}}(\mathbf{y})} \text{ si } p_{\mathbf{Y}}(\mathbf{y}) \neq 0,\ p(\theta|\mathbf{y}) = 1 \text{ sinon.} \tag{7}$$

Dans un tel cadre, des hypothèses classiques de régularité de la vraisemblance par rapport à $\theta$ (voir par ex. Lehmann (1999)) permettent à la matrice d'information de Fisher $\mathcal{I}(\theta) = (\mathcal{I}(\theta)_{i,j})_{i,j=1}^d$ d'être bien définie selon :

$$\mathcal{I}(\theta)_{i,j} = -\int_{\mathcal{Y}} [\partial^2_{\theta_i \theta_j} \log \ell(y|\theta)]\, \ell(y|\theta) d\mu(y). \tag{8}$$

Notons enfin que le théorème d'extension de Kolmogorov propose un cadre qui inclut le travail qui précède et qui l'étend à toute valeur de $k$ : il existe un espace probabilisé $(\overline{\Omega}, \overline{\mathscr{P}}, \overline{\mathbb{P}})$ tel que pour tout $k$, et tout $A \in \sigma_T$, $B_1 \in \sigma_{Y_1}, \ldots, B_k \in \sigma_{Y_k}$,

$$\overline{\mathbb{P}}(A, B_1, \ldots, B_k) = \mathbb{P}(A, B_1, \ldots, B_k) = \int_{T(A)} \mathbb{P}^{\otimes k}_{\mathbf{Y}|T=\theta}(\mathbf{Y}(B_1 \times \cdots \times B_k)) d\pi(\theta). \tag{9}$$

# 3 Prior de référence sous contraintes

## 3.1 Information mutuelle et prior de référence généralisé

Cette communication prend pour support la théorie étendue des priors de références telle que proposée et décrite dans (Van Biesbroeck, 2023). Dans notre travail, nous considérons l'information mutuelle $I_{D_f}$, définie à partir d'une $f$-divergence $D_f$ comme mesure de dissimilarité :

$$I_{D_f}(\pi|k) = \int_\Theta D_f(\mathbb{P}_{\mathbf{Y}|T=\theta}||\mathbb{P}_{\mathbf{Y}}) d\pi(\theta), \tag{10}$$

$$\text{avec} \quad D_f(\mathbb{P}_{\mathbf{Y}|T=\theta}||\mathbb{P}_{\mathbf{Y}}) = \int_{\mathcal{Y}^k} f\left(\frac{\ell_k(\mathbf{y}|\theta)}{p_{\mathbf{Y}}(\mathbf{y})}\right) d\mu^{\otimes k}(\mathbf{y}). \tag{11}$$

Le cadre des $f$-divergences laisse plutôt ouvert le choix de fonction $f$. A titre d'exemple, si $f = -\log$, alors $D_f$ n'est autre que la célèbre divergence de Kullback-Leibler. Dans notre travail, nous nous limitons à des fonctions $f$ mesurables, localement bornées et dont les comportements asymptotiques en 0 et en $+\infty$ sont contrôlés selon :

$$f(x) \underset{x\to 0^+}{=} \alpha x^\beta + o(x^\beta), \quad f(x) \underset{x\to\infty}{=} O(x), \tag{12}$$

avec $\alpha < 0$, $\beta \in ]0, 1[$.

Nous supposons également $\Theta$ compact. Dans ce cadre, nous adaptons et ré-exprimons ci-dessous la définition du prior de référence généralisé sous la considération de la mesure de dissimilarité $D_f$.

*Définition* 1 ($D_f$-prior de référence (Van Biesbroeck, 2023)). Soit $\mathcal{P}$ une classe de priors sur $\Theta$. Un prior $\pi^* \in \mathcal{P}$ est appelé $D_f$-prior de référence sur la classe $\mathcal{P}$ au taux $\varphi(k)$ si

$$\forall \pi \in \mathcal{P}, \lim_{k\to\infty} \varphi(k)(I_{D_f}(\pi^*|k) - I_{D_f}(\pi|k)) \geq 0. \tag{13}$$

Le théorème ci-après est extrait de (Van Biesbroeck, 2023, Theorem 1). Sous des hypothèses formelle que l'auteur décrit, il donne une expression de la limite impliquée dans l'équation (13).

**Théorème 1** (Van Biesbroeck (2023)). *Soit $\pi$ un prior qui admet une densité continue par rapport à la mesure de Lebesgue sur $\Theta$, que l'on note également $\pi$ ici sans ambiguïté. Alors,*

$$\lim_{k\to\infty} k^{d\beta/2} I_{D_f}(\pi|k) = l(\pi) = \alpha C_\beta \int_\Theta \pi(\theta)^{1+\beta} |\mathcal{I}(\theta)|^{-\beta/2} d\theta \tag{14}$$

*avec* $C_\beta = (2\pi)^{d\beta/2}(1-\beta)^{-d/2}$.

Le choix de la classe $\mathcal{P}$ dans la définition 1 reste ouvert, et bien que le théorème ci-dessus amène à se restreindre à la classe —encore large— des priors à densités continues, des restrictions additionnelles pourraient très bien y être ajoutées. C'est cette idée que nous explorons dans la section qui suit, avec l'introduction de contraintes linéaires sur le prior.

## 3.2 Résultat principal

La principale contribution de ce travail constitue le théorème qui suit, en proposant une expression d'un prior de référence sous un certain nombre de contraintes linéaires.

Le cadre de travail général se limite ici à celui des priors qui admettent une densité continue par rapport à la mesure de Lebesgue. A ce titre et à dessein de simplification, la lettre $\pi$ servira dorénavant plutôt la désignation générique d'une densité plutôt que d'un prior. $E$ désigne alors l'espace des fonctions continues de $\Theta$ dans $\mathbb{R}$, il est muni de la norme infinie : $\|p\| = \sup_\Theta |p|$ pour tout $p \in E$.

*Hypothèse* 1. Une famille de fonctions mesurables $g_1, \dots, g_p : \Theta \longrightarrow \mathbb{R}$ est dite satisfaisant l'hypothèse 1 si elles sont intégrables sur $\Theta$ et que les fonctions $g_0, \dots, g_p$ sont linéairement indépendantes, en définissant $g_0 : \theta \mapsto 1$.

**Théorème 2.** *Soient $g_1, \dots, g_p$ des applications mesurables de $\Theta$ vers $\mathbb{R}$ qui satisfont l'hypothèse 1. On définit $\mathcal{P}$ comme la classe des priors admettant une densité $\pi \in E$ positive et telle que $\forall j = 1, \dots, p$, $\int_\Theta g_i(\theta)\pi(\theta)d\theta = c_j$ pour des certains $c_j \in \mathbb{R}$. Si $\mathcal{P}$ est non vide, alors il existe un unique $D_f$-prior de référence sur $\mathcal{P}$. S'il est strictement positif sa densité $\pi^*$ s'écrit*

$$\pi^* = J \cdot (\lambda_0 + \sum_{i=1}^{p} \lambda_i g_i)^{1/\beta}, \tag{15}$$

*pour certains scalaires $\lambda_0, \dots, \lambda_p \in \mathbb{R}$. Réciproquement, un prior de $\mathcal{P}$ dont la densité s'exprime sous la forme ci-dessus est l'unique $D_f$-prior de référence. $J$ désigne la densité du prior de Jeffreys : $J(\theta) = |\mathcal{I}(\theta)|^{1/2} / \int_\Theta |\mathcal{I}(\tilde{\theta})|^{1/2} d\tilde{\theta}$.*

Les contraintes linéaires auxquelles se limitent ce théorème sauraient intuitivement servir de contraintes de moment. Un exemple simple pourrait être de choisir $p = d$ et $g_i = \theta_i$ pour fixer les espérances a priori. Par ailleurs, ce format de contraintes peut aussi s'appliquer aux paramètres d'ellicitation prédictive (Bousquet, 2023) : soit des percentiles prédictifs $(t_{\delta_i}, \delta_i)_i$, on contraint le prior $\pi$ à satisfaire :

$$\mathbb{P}(h(\mathbf{Y}) \leq t_{\delta_i}) = \int_\Theta \int_{\mathcal{Y}^k} \mathbb{1}_{h(\mathbf{y}) \leq t_{\delta_i}} \ell_k(\mathbf{y}|\theta) d\mu^{\otimes k}(\mathbf{y}) \pi(\theta) d\theta \tag{16}$$

pour une certaine fonction $h$.

## 3.3 Démonstration

Rappelons la notation $l$ issue de l'équation (14) :

$$l(\pi) = \alpha C_\beta \int_\Theta \pi(\theta)^{1+\beta} |\mathcal{I}(\theta)|^{-\beta/2} d\theta. \tag{17}$$

Le théorème 1 fait le lien entre l'optimisation de $l$ et le $D_f$-prior de référence. Ce dernier est en effet le point en lequel $l$ atteint son maximum.

Remarquons que $\Theta$ étant compact, la paire $(E, \|\cdot\|)$ constitue un espace de Banach dont la restriction $U$ composée des fonctions strictement positives est un sous-ensemble ouvert et convexe, sur lequel $l$ définit une fonction concave à valeurs dans $\mathbb{R}$.

Différentions $l$ sur $U$. Pour ceci on écrit $l = \phi_1 \circ \phi_2$ avec

$$\phi_1 : \pi \in E \longmapsto \pi^{1+\beta} \in E; \qquad \phi_2 : \pi \in E \longmapsto \alpha C_\beta \int_\Theta \pi(\theta)|\mathcal{I}(\theta)|^{-\beta/2} d\theta. \tag{18}$$

Comme $\phi_2$ est une application linéaire continue de $E$ dans $\mathbb{R}$, $l$ est différentiable tant que $\phi_1$ l'est, avec :

$$dl(\pi) = d\phi_1(\phi_2(\pi)) \circ \phi_2. \tag{19}$$

Soit $\pi \in U$. Pour tout $\varepsilon > 0$ il existe $\tilde{\varepsilon} > 0$ tel que tant que $|x| < \|\pi\|$ et $|u| < \tilde{\varepsilon}$ alors $|(x+u)^{1+\beta} - x^\beta - (1+\beta)x^\beta u| < \varepsilon|u|$. Ainsi, pour tout $h \in E$ tel que $\|h\| < \tilde{\varepsilon}$, on peut écrire

$$\|\phi_1(\pi + h) - \phi_1(\pi) - (1+\beta)\pi^\beta h\| < \varepsilon\|h\|. \tag{20}$$

On conclut que $\phi_1$ est différentiable sur $U$ avec $d\phi_1(\pi)h = (1+\beta)\pi^\beta h$, pour tout $\pi \in U$, $h \in E$. Cette différentielle est de plus continue, dont on déduit que $l$ est également continûment différentiable.

Ainsi, considérant la contrainte additionnelle $\int_\Theta \pi(\theta)d\theta = 1$, ce problème peut être traité en appliquant le théorème des extrema liés (cf. par ex. Cartan (2007)) selon lequel il existe $\lambda_0, \ldots, \lambda_p \in \mathbb{R}^{p+1}$ tels que tout extremum $\pi^* \in U \cap C$ (où $C$ désigne les fonctions qui satisfont les contraintes) de $l$ et satisfaisant les contraintes vérifie

$$dl(\pi^*)h - \lambda_0 \int_\Theta h(\theta)d\theta - \sum_{i=1}^{p} \lambda_i \int_\Theta h(\theta)g_i(\theta)d\theta = 0 \tag{21}$$

pour tout $h \in E$. Enfin, comme $dl(\pi)h = \alpha C_\beta(1+\beta)\int_\Theta \pi(\theta)^\beta |\mathcal{I}(\theta)|^{-\beta/2} h(\theta)d\theta$, on obtient, quitte à renommer les $\lambda_i$,

$$\pi^*(\theta) = J(\theta)\left(\lambda_0 + \sum_{i=1}^{p} \lambda_i g_i(\theta)\right)^{1/\beta}. \tag{22}$$

Aussi, la stricte concavité de $l$ implique que, (i) s'il existe, $\pi^*$ est l'unique argument maximal de $l$ sur $U \cap C$, et que (ii) si réciproquement un prior $\pi^* \in U \cap C$ satisfait l'équation (22) alors il est l'unique l'argument maximal de $l$ sur $U \cap C$.

Pour conclure, on remarque que les densités des priors positifs satisfaisant les contraintes constituent la fermeture de l'ensemble $U \cap C$. Puisque $\pi^*$ tel que défini ci-dessus maximise $l$ sur $U \cap C$, la continuité de $l$ sur sa fermeture induit le caractère maximal de $\pi^*$ sur celle-ci également. Autrement dit, le prior auquel il est associé est le $D_f$-prior de référence sur $\mathcal{P}$.

# 4 Conclusion

Le question du choix du prior en inférence bayésienne reste ouverte et complexe. D'une part, le point de vue de la théorie des priors de référence suggère la minimisation de l'influence de toute information a priori ; d'autre part, de nombreuses applications pratiques plébiscitent l'emploi du canevas bayésien pour l'introduction d'un jugement a priori.

Notre communication propose une réconcilitation de ces deux mondes en définissant un cadre sur lequel leur intersection est possible. Notre prior de référence se construit ici sous contraintes, et explicite en notre sens le bon cadre pour l'introduction d'information a priori.

Ce dernier formalisme s'appuie sur une définition plus étendue de la théorie des priors de référence qui propose la considération de mesures de dissimilarité plus large, en complément à l'historique divergence de Kullback-Leibler. La souplesse de leur cadre et de leur résultat est appuyé par notre travail et ses ouvertures.

# Références


Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B*, 41(2), 113–147. doi:10.1111/j.2517-6161.1979.tb01066.x.

Bousquet, N. (2023). Discussion of "specifying prior distributions in reliability applications" : Towards new formal rules for informative prior elicitation ? *Applied Stochastic Models in Business and Industry*. doi:10.1002/asmb.2794.

Cartan, H. (2007). *Cours de calcul différentiel*. Sciences et techniques. Hermann, 2ème édition.

Clarke, B. S. & Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1), 37–60. doi:10.1016/0378-3758(94)90153-8.

Gauchy, C., Van Biesbroeck, A., Feau, C., & Garnier, J. (2023). Inférence variationelle de lois a priori de référence. Dans *Proceedings des 54èmes Journées de Statistiques de la SFDS (JdS)*.

Gu, M., Wang, X., & Berger, J. O. (2018). Robust Gaussian stochastic process emulation. *The Annals of Statistics*, 46(6A), 3038–3066. doi:10.1214/17-AOS1648.

Keefe, M. J., Ferreira, M. A. R., & Franck, C. T. (2019). Objective Bayesian Analysis for Gaussian Hierarchical Models with Intrinsic Conditional Autoregressive Priors. *Bayesian Analysis*, 14(1), 181–209. doi:10.1214/18-BA1107.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer Texts in Statistics. Springer New York, NY, 1 édition.

Muré, J. (2021). Propriety of the reference posterior distribution in Gaussian process modeling. *The Annals of Statistics*, 49(4), 2356–2377. doi:10.1214/20-AOS2040.

Nalisnick, E. & Smyth, P. (2017). Learning approximately objective priors. Dans *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Van Biesbroeck, A. (2023). Generalized mutual information and their reference prior under Csizar f-divergence. arXiv.2310.10530. doi:10.48550/arXiv.2310.10530.

# Inférence distribuée pour les modèles de mélange de processus de Dirichlet dans l'apprentissage fédéré

Reda Khoufache [1] & Mustapha Lebbah [2] & Hanene Azzag [3] & Etienne Goffinet [4] & Djamel Bouchaffra[5]

[1] *DAVID, UVSQ, Université Paris-Saclay, Versailles. Email: reda.khoufache@uvsq.fr*
[2] *DAVID, UVSQ, Université Paris-Saclay, Versailles. Email: mustapha.lebbah@uvsq.fr*
[3] *LIPN, Université Sorbonne Paris Nord, France. Email: azzag@univ-paris13.fr*
[4] *Technology Innovation Institute, Abu Dhabi, UAE. Email: etienne.goffinet@tii.ae*
[5] *CDTA, Algérie. Email: djamel.bouchaffra@gmail.com*

**Résumé.** Cet article a déjà été publié dans [16]. Il présente une méthode d'inférence distribuée (DisCGS) pour les modèles de mélange de processus de Dirichlet (DPMM). Le DPMM est beaucoup utilisé pour résoudre les problèmes de (*clustering*), offrant l'avantage d'estimer automatiquement le nombre de *clusters* durant l'inférence via la modélisation bayésienne non paramétrique. Cependant, leur processus d'inférence est considérablement lent lorsque le nombre d'observations est grand. Notre approche, basée sur l'échantillonneur de Gibbs qui est une méthode de Monte-Carlo par chaînes Markov (MCMC), est conçue pour être exécutée sur un environnement distribué, notamment dans le contexte de l'apprentissage fédéré horizontal. La méthode DisCGS a montré des performances remarquables. Par exemple, pour 100 000 observations, notre approche atteint 100 itérations en seulement 3 minutes, soit un facteur de réduction du temps d'exécution de 200 par rapport à l'algorithme centralisé qui nécessite environ 12 heures. Le code source est accessible publiquement sur `https://github.com/redakhoufache/DisCGS`.

**Mots-clés.** Apprentissage fédéré, Calcul distribué, Modèles de mélange de processus de Dirichlet, Monté-Carlo par chaînes de Markov, Modélisation bayésienne non-paramétrique

**Abstract.** This paper has already been published in [16]. It introduces DisCGS, a distributed Markov Chain Monte Carlo (MCMC) inference method for Dirichlet Process Mixture Models (DPMMs) using sufficient statistics. Our method uses collapsed Gibbs sampling and is designed to work on distributed data across independent and heterogeneous machines, making it suitable for horizontal federated learning. Our approach demonstrates promising results and remarkable scalability. For instance, the centralized algorithm requires approximately 12 hours to complete 100 iterations while our approach achieves the same number of iterations in just 3 minutes, reducing the execution time by a factor of 200 without compromising clustering performance. The source code is publicly available at `https://github.com/redakhoufache/DisCGS`.

**Keywords.** Federated learning, Distributed computing, Dirichlet process mixture models, Markov Chain Monte Carlo, Bayesian non-parametric

# 1 Introduction

Les modèles de mélange de processus de Dirichlet (DPMM) représentent une extension des modèles de mélange vers une approche bayésienne non paramétrique. Ces modèles probabilistes génératifs supposent que les observations suivent une loi de mélange avec un nombre infini de composantes, dont les paramètres sont aléatoires et suivent une loi a priori. Ils sont largement utilisés pour traiter les problèmes de *clustering*, en particulier lorsque le nombre de *clusters* est inconnu car ils ont la capacité de l'estimer durant le processus d'inférence. L'algorithme de Gibbs [15] est une méthode MCMC qui infère les paramètres du DPMM en échantillonnant la variable latente de chaque observation selon la loi a posteriori. Bien que le DPMM offre l'avantage de découvrir de nouvelles structures et d'estimer automatiquement le nombre de *clusters*, l'étape d'inférence devient considérablement lente et ne passe pas à l'échelle lorsque le nombre d'observations est très grand. Ce qui limite son usage pour le traitement des données massives et son applicabilité dans des cas d'usages réels. Le calcul distribué consiste à distribuer les données sur des *workers*, ce qui permet d'effectuer des opérations parallèles, accélérant ainsi les calculs. L'apprentissage fédéré (FL), introduit dans [10], a révolutionné l'apprentissage distribué à grande échelle en entraînant les modèles directement sur des dispositifs locaux, que nous appelons "workers". L'une des caractéristiques du FL est que les données peuvent avoir des distributions différentes d'un *worker* à l'autre. Cette hétérogénéité nécessite des méthodes de *clustering* robustes. Par ailleurs, la nature décentralisée du FL pose des défis statistiques et informatiques, et la complexité des problèmes de *clustering* augmente considérablement dans ce contexte.

Les problèmes de *clustering* dans le contexte fédéré ont suscité un grand intérêt récemment. [11] a introduit une approche fédérée du K-Means, [19] a proposé un algorithme Fuzzy $c$-Means fédéré, tandis que [9] a présenté une méthode de *clustering* fédérée basée sur des modèles probabilistes. Plusieurs approches parallèles et distribuées d'inférence du DPMM ont été proposées dans la littérature. [12] a introduit un reparamétrage du processus de Dirichlet pour l'apprentissage des *clusters*, tandis que [22] a intégré des variables auxiliaires pour permettre la parallélisation. Cependant, [5] a montré que ces approches sont impraticables en raison de distributions déséquilibrées. Une autre méthode MCMC parallélisée a été présentée dans [1], combinant un échantillonneur de Gibbs restreint avec des propositions de division/fusion pour garantir l'ergodicité. Dans [3], cette approche a été étendue à un environnement distribué. [6] a introduit un algorithme d'inférence distribuée pour le DPMM basé sur l'échantillonnage par tranche. Une méthode d'estimation distribuée pour le DPMM est présentée dans [21], qui crée de nouvelles composantes localement au niveau des *workers*, suivies d'un schéma de consolidation probabiliste au niveau du *master*. [13] propose une approche d'inférence distribuée basée sur la méthode MCMC, utilisant un échantillonneur de Gibbs au niveau des *workers* pour découvrir de nouveaux *clusters*, et un autre échantillonneur sur les moyennes de chaque *cluster* au niveau du *master* pour l'agrégation des résultats. Cette méthode suppose des variances connues, limitant ainsi son applicabilité.

Dans cet article, nous présentons une nouvelle approche distribuée, DisCGS, basée sur les méthodes MCMC pour l'inférence du DPMM, en utilisant des statistiques suffisantes. Il est important de noter que notre approche se concentre sur un algorithme d'inférence spécifique pour le DPMM, le Gibbs Sampler proposé dans [15].

## 2 Modèle de mélange de processus de Dirichlet

Soient $n$ et $d$ deux entiers naturels, $\mathbf{x} = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n \times d}$ le jeu de données, où $(\cdot)^T$ est l'opérateur de transposition. Soit $\mathbf{z} = (z_1, \cdots, z_n)$ le vecteur d'appartenance (la partition), où $z_i$ est une variable latente telle que $z_i = k$ signifie que l'observation $x_i$ appartient au *cluster* $k$. le DPMM suppose que les observations sont générées suivant ce modèle :

$$\begin{aligned} x_i \mid \{z_i = k, \theta_k\} &\overset{\text{i.i.d.}}{\sim} f(x_i, \theta_k), \forall i \in \{1, \cdots, n\} \\ \theta_k &\overset{\text{i.i.d.}}{\sim} G_0, \forall k \in \{1, 2, \cdots\} \\ z_i &\overset{\text{i.i.d.}}{\sim} \text{Mult}(\pi), \forall i \in \{1, \cdots, n\} \\ \pi &\sim SB(\alpha), \end{aligned}$$

Sous cette hypothèse, une observation $x_i$ est générée en échantillonnant d'abord $z_i$ suivant la loi multinoulli de paramètre $\pi = (\pi_k)_{k=1}^{\infty}$. Ensuite, $x_i$ est générée suivant $f(x_i, \theta_{z_i})$, où $f(\cdot, \theta_k)$ est la densité associée au *cluster* $k$ dont le paramètre $\theta_k$ suit une loi a priori $G_0$ (distribution de base); dans le cas Gaussien multivarié, nous avons $\theta_k = (\mu_k, \Sigma_k)$ avec $\mu_k \in \mathbb{R}^d$ et $\Sigma_k \in \mathbb{R}^{d \times d}$ une matrice symétrique semi-définie positive. Le vecteur $\pi$ suit le processus du *Stick-Breaking* [17] de paramètre de concentration $\alpha > 0$. Dans ce qui suit, nous supposons que $f$ est une distribution gaussienne multivariée et que $G_0$ est l'a priori conjugué Normal Inverse Wishart [14] (NIW) de paramètres $(\mu_0, \kappa_0, \Psi_0, \nu_0)$. Nous allons noter $\Theta = \{\theta_k,\ k > 1\}$ l'ensemble des paramètres et $\Omega = (\alpha, G_0)$ l'ensemble des hyperparamètres.

**Inférence.** L'algorithme de Gibbs [15] permet d'inférer les paramètres via la méthode de Monte-Carlo. Il alterne entre la mise à jour de la partition $\mathbf{z}$ et celle des paramètres associés à chaque *cluster*. Pendant la première étape, l'échantillonneur de Gibbs simule la loi jointe a posteriori $\text{p}(\mathbf{z} \mid \mathbf{x}, \Theta, \Omega)$ en tirant chaque $z_i$ suivant la loi conditionnelle $\text{p}(z_i \mid \mathbf{x}, \mathbf{z}_{-i}, \Theta, \Omega)$, où $\mathbf{z}_{-i} = \{z_l,\ l \neq i\}$. Durant la seconde étape, les paramètres associés au *cluster* $k$ sont tirés selon la loi a posteriori $\text{p}(\theta_k | \mathbf{x}_k, G_0)$, avec $\mathbf{x}_k = \{x_i,\ z_i = k\}$ le contenu du *cluster* $k$. Le calcul de cette distribution peut se faire analytiquement comme $G_0$ est la loi a priori conjuguée pour la densité $f$. Également, il est possible d'intégrer les paramètres $\theta_k$, ce qui permet de sauter l'étape de mise à jour des paramètres, ainsi seuls les variables latentes $z_i$ sont mises à jour. Ceci nous amène à une variante de l'algorithme de Gibbs qui correspond au 3 ème algorithme proposé dans [15].

## 3 La méthode proposée

L'objectif de notre approche consiste à distribuer le processus d'inférence sans compromettre les performances de l'estimateur initial. Pour ce faire, après avoir distribué uniformément les données entre les *workers*, notre algorithme alterne entre des phases d'inférence aux niveaux *master* et *worker*.

## 3.1 DisCGS au niveau *worker*

Soit $\mathbf{x}^j = \{x_1^j, \cdots, x_{n^j}^j\}$ l'ensemble d'observations assignées au $j$-ième *worker*, $n^j$ le cardinal de $\mathbf{x}^j$, et $\mathbf{z}^j = (z_1^j, \cdots z_{n^j}^j)$ la partition locale où $z_i^j$ est une variable latente locale de sorte que $z_i^j = k$ signifie que l'observation $x_i^j$ (assignée au *worker* $j$) appartient au *cluster* local $k$. Les variables latentes locales sont mises à jour en utilisant la variante de l'échantillonneur de Gibbs introduit dans la section 2. Chaque $z_i^j$ est échantillonné selon $\mathrm{p}(z_i^j \mid \mathbf{z}_{-i}^j, \mathbf{x}^j, \Omega) \propto$

$$\begin{cases} n_k^j \mathrm{p}(x_i^j \mid z_i^j = k, \mathbf{x}_k^j, G_0), & \textit{cluster} \text{ existant } k, \quad (1) \\ \alpha \mathrm{p}(x_i^j \mid \Omega), & \text{nouveau } \textit{cluster}, \quad (2) \end{cases}$$

où $n_k^j$ et $\mathbf{x}_k^j$ sont respectivement la taille et le contenu du *cluster* $k$ du *worker* $j$. Sous l'hypothèse de conjugaison, les distributions prédictives a posteriori et a priori (équations 1 et 2 respectivement) sont calculées analytiquement [14]. Après avoir mis à jour la partition, nous calculons les statistiques suffisantes [18] associées à chaque *cluster*. Dans le cas Gaussien multivarié, les statistiques suffisantes $(T_k^j, S_k^j)$ pour un *cluster* $\mathbf{x}_k^j$ sont données par [7]:

$$T_k^j = \frac{1}{n_k^j} \sum_{x \in \mathbf{x}_k^j} x \in \mathbb{R}^d, \quad (3)$$

$$S_k^j = \sum_{x \in \mathbf{x}_k^j} (x - T_k^j)(x - T_k^j)^T \in \mathbb{R}^{d \times d}. \quad (4)$$

Enfin, les statistiques suffisantes et les tailles de chaque *cluster* sont envoyées au *master*.

## 3.2 DisCGS au niveau *master*

Le *master* reçoit de chaque *worker* les tailles et les statistiques suffisantes associées à chaque *cluster*. L'objectif est d'estimer le vecteur d'appartenance global $\mathbf{z} = (z_1, \cdots, z_n)$ et de mettre à jour les hyperparamètres associés à chaque *cluster*. À ce niveau, au lieu d'assigner les observations une par une à leur *cluster* global, nous assignons un groupe d'observations qui partagent déjà le même *cluster* local (c'est-à-dire au niveau du *worker*) à un *cluster* global au niveau du *master*. Ainsi, les observations assignées au même *cluster* global auront la même étiquette. Nous échantillonnons la variable latente globale $z_h^j$ du *cluster* $\mathbf{x}_h^j$ (*cluster* local $h$ du *worker* $j$) suivant $\mathrm{p}(z_h^j \mid \mathbf{z}_{-h}^{-j}, \mathbf{x}, \Omega) \propto$

$$\begin{cases} n_k \mathrm{p}(\mathbf{x}_h^j \mid z_h^j = k, \mathbf{x}_k, G_0), & \textit{cluster} \text{ existant } k, \quad (5) \\ \alpha \mathrm{p}(\mathbf{x}_h^j \mid G_0), & \text{nouveau } \textit{cluster}. \quad (6) \end{cases}$$

Les distributions prédictives a posteriori et a priori jointes (équations 5 et 6 respectivement) sont calculées analytiquement en utilisant uniquement les statistiques suffisantes, c'est-à-dire sans avoir accès au contenu du *cluster* $\mathbf{x}_h^j$. En effet, nous avons:

$$\mathrm{p}\left(\mathbf{x}_h^j \mid \Omega\right) = \pi^{-n_h^j \frac{d}{2}} \cdot \frac{\kappa_0^{d/2}}{\left(\kappa_h^j\right)^{d/2}} \cdot \frac{\Gamma_d\left(\nu_h^j/2\right)}{\Gamma_d\left(\nu_0/2\right)} \cdot \frac{\left|\Psi_0\right|^{\nu_0/2}}{\left|\Psi_h^j\right|^{\nu_h^j/2}}$$

où $|\cdot|$ est le déterminant, et les hyperparamètres $(\mu_h^j, \kappa_h^j, \Psi_h^j, \nu_h^j)$ sont obtenus comme suit:

$$\mu_h^j = \frac{\kappa_0\mu_0 + n_h^j T_h^j}{\kappa_h^j}, \quad \kappa_h^j = \kappa_0 + n_h^j, \quad \nu_h^j = \nu_0 + n_h^j,$$
$$\Psi_h^j = \Psi_0 + S_h^j + \frac{\kappa_0 n_h^j}{\kappa_h^j}\left(\mu_0 - T_h^j\right)\left(\mu_0 - T_h^j\right)^T,$$

où $T_h^j$ et $S_h^j$ sont les statistiques suffisantes renvoyées par les *workers*. Par ailleurs, on a:

$$\mathrm{p}(\mathbf{x}_h^j \mid z_h^j = k, \mathbf{x}_k, G_0) = \pi^{\frac{-dn_h^j}{2}} \cdot \frac{\kappa_k^{d/2}}{\left(\kappa_h^j\right)^{d/2}} \cdot \frac{\Gamma_d\left(\nu_h^j/2\right)}{\Gamma_d\left(\nu_k/2\right)} \cdot \frac{|\Psi_k|^{\nu_k/2}}{\left|\Psi_h^j\right|^{\nu_h^j/2}}$$

où les paramètres de la loi a posteriori $(\mu_k, \kappa_k, \Psi_k, \nu_k)$ associés au *cluster* global $k$, sont mis à jour à partir de la loi a priori ainsi:

$$\mu_k = \frac{\kappa_0\mu_0 + n_k T_k}{\kappa_k}, \quad \kappa_k = \kappa_0 + n_k, \quad \nu_k = \nu_0 + n_k,$$
$$\Psi_k = \Psi_0 + S_k + \frac{\kappa_0 n_k}{\kappa_k}\left(\mu_0 - T_k\right)\left(\mu_0 - T_k\right)^T,$$

avec $T_k$ et $S_k$, les statistiques suffisantes agrégées lorsque des *clusters* locaux sont assignées au même *cluster* global $k$ sont calculées comme suit:

$$T_k = \frac{1}{n_k} \sum_{j,h|\,\mathbf{z_h^j}=\mathbf{k}} n_h^j \cdot T_h^j,$$
$$S_k = \sum_{j,h|\,\mathbf{z_h^j}=\mathbf{k}} S_h^j + \sum_{j,h|\,\mathbf{z_h^j}=\mathbf{k}} \left(n_j^h \cdot T_h^j \cdot T_h^{j^T}\right) - n_k \cdot T_k \cdot T_k^T.$$

Le pseudo-code du processus d'inférence au niveau *worker* et *master* sont donnés dans [16].

## 3.3 L'échantillonneur de Gibbs dans l'apprentissage fédéré

Dans le contexte fédéré, les observations se retrouvent sur différents composants (*workers*). Cela correspond à la décomposition horizontale des données. Chaque *worker* est initialisé avec le même modèle global. Ensuite, chaque *worker* met à jour son modèle local en utilisant ses données privées via l'échantillonneur de Gibbs détaillé dans la section 3.1; cette étape permet de découvrir les *clusters* locaux et d'inférer le DPMM local. Ensuite, les statistiques suffisantes et les tailles associées à chaque *cluster* local sont calculées et transmises au serveur (*master*). Ce dernier procède à la mise à jour du modèle global et à l'estimation de la partition globale sans avoir accès aux données. Ce processus est réalisé en utilisant l'échantillonneur de Gibbs décrit dans la section 3.2. Le modèle à jour est ensuite partagé avec chaque composant. Ce processus itératif se poursuit en alternant ces deux étapes jusqu'à ce que le modèle global soit estimé.

| Jeu de données | $n$ | $d$ | $K$ | Description |
|---|---|---|---|---|
| Synthetic 10K | 10000 | 2 | 6 | Jeu de donnnées synthetique généré selon des composantes gaussiennes de dimension 2. |
| Mnist | 70000 | 8 | 10 | Chiffres écrits à la main, Lecun et al. |
| Fashion mnist | 70000 | 8 | 10 | Images d'articles Zalando [23]. |
| Balanced | 131600 | 8 | 47 | Chiffres et lettres écrits à la main [2]. |
| Digits | 280000 | 8 | 10 | Chiffres écrits à la main [2]. |
| UrbanGB | 360177 | 3 | 469 | Coordonnées (longitude et latitude) d'accidents de la route [4]. |

Table 1: Description des jeux de données utilisés pour évaluer les performances *clustering* de notre approche. $n$ est le nombre d'observations, $d$ la dimension, $K$ le nombre de *clusters*.

# 4 Expériences

Pour évaluer notre approche, nous présentons trois expériences sur des jeux de données synthétiques et réels: d'abord un *benchmark* comparant les performances de *clustering* et la vitesse de convergence, puis une comparaison des temps d'exécution de l'algorithme centralisé et distribué, et enfin, une expérience qui évalue le passage à l'échelle de notre méthode.

## 4.1 Implémentation et environnement distribué

Dans ce qui suit, nous utilisons la loi a priori non informative pour les algorithmes CGS et DisCGS. Ainsi, nous posons $\mu_0$ et la matrice de précision $\Psi_0$ égaux au vecteur moyen et à la matrice de covariance des données. $\kappa_0$ et $\nu_0$ représentent notre confiance en $\mu_0$ et $\Psi_0$, et sont fixés à leurs plus petites valeurs, qui sont 1 et $d+1$. L'état initial est une partition à un *cluster*. Nous avons exécuté les algorithmes en utilisant la machine Neowise (1 CPU AMD EPYC 7642, 48 cores/CPU) hébergée par le *cluster* grid5000 [1].

## 4.2 Performances de *clustering*

Dans cette expérience, nous comparons DisCGS avec deux algorithmes distribués: M-R [6] et SubC [3], ainsi qu'avec deux algorithmes parallélisés: Kmeans et GMM. Toutes les exécutions sont effectuées en distribuant les données sur 32 cœurs. Nous utilisons 6 jeux de données décrits dans le tableau 1. Les jeux de données d'images sont encodés en un vecteur de dimension 8 en utilisant un auto-encodeur variationnel. Pour évaluer les performances de *clustering*, nous calculons les trois métriques de *clustering*; l'indice de Rand ajusté (ARI) [8], l'information mutuelle normalisée (NMI) [20], et l'*accuracy* (ACC) [24].

La table 2 présente la moyenne et l'écart type des trois métriques obtenues par chaque méthode sur 10 essais. Les résultats montrent que notre méthode proposée surpasse les autres méthodes ou obtient le deuxième meilleur score sur presque tous les jeux de données. Il est important de noter que dans cette expérience, nous nous concentrons uniquement sur la comparaison des performances de *clustering*; nous ne comparons pas les temps d'exécution car les autres approches proposent une inférence différente de l'échantillonneur de Gibbs.

[1] `https://www.grid5000.fr/w/Grid5000:Home`

| Jeu de données | | DisCGS | M-R | SubC | GMM | Kmeans |
|---|---|---|---|---|---|---|
| Synthetic 10K | ARI | **0.80** ∓ 0.06 | 0.10 ∓ 0.01 | 0.38 ∓ 0.07 | 0.80 ∓ 0.07 | 0.75 ∓ 0.03 |
| | NMI | 0.86 ∓ 0.04 | 0.12 ∓ 0.01 | 0.61 ∓ 0.08 | **0.88** ∓ 0.04 | 0.85 ∓ 0.01 |
| | ACC | **0.88** ∓ 0.06 | 0.29 ∓ 0.01 | 0.38 ∓ 0.08 | 0.85 ∓ 0.08 | 0.76 ∓ 0.05 |
| Mnist | ARI | **0.72** ∓ 0.01 | 0.20 ∓ 0.07 | 0.66 ∓ 0.04 | 0.39 ∓ 0.02 | 0.26 ∓ 0.01 |
| | NMI | 0.74 ∓ 0.00 | 0.38 ∓ 0.08 | **0.79** ∓ 0.01 | 0.69 ∓ 0.01 | 0.62 ∓ 0.00 |
| | ACC | **0.79** ∓ 0.01 | 0.30 ∓ 0.06 | 0.71 ∓ 0.03 | 0.38 ∓ 0.02 | 0.23 ∓ 0.01 |
| Fashion-Mnist | ARI | **0.45** ∓ 0.02 | 0.35 ∓ 0.02 | 0.40 ∓ 0.02 | 0.41 ∓ 0.02 | 0.37 ∓ 0.02 |
| | NMI | **0.60** ∓ 0.01 | 0.54 ∓ 0.01 | **0.60** ∓ 0.01 | 0.59 ∓ 0.02 | 0.57 ∓ 0.01 |
| | ACC | **0.55** ∓ 0.02 | 0.45 ∓ 0.03 | 0.48 ∓ 0.01 | 0.52 ∓ 0.04 | 0.48 ∓ 0.01 |
| Balanced | ARI | **0.35** ∓ 0.00 | 0.02 ∓ 0.01 | 0.05 ∓ 0.02 | 0.35 ∓ 0.01 | 0.22 ∓ 0.00 |
| | NMI | 0.59 ∓ 0.00 | 0.17 ∓ 0.06 | 0.31 ∓ 0.04 | **0.63** ∓ 0.00 | 0.54 ∓ 0.00 |
| | ACC | **0.46** ∓ 0.01 | 0.07 ∓ 0.02 | 0.10 ∓ 0.02 | 0.44 ∓ 0.02 | 0.30 ∓ 0.01 |
| Digits | ARI | 0.55 ∓ 0.01 | 0.28 ∓ 0.14 | **0.74** ∓ 0.05 | 0.46 ∓ 0.02 | 0.36 ∓ 0.01 |
| | NMI | 0.71 ∓ 0.01 | 0.51 ∓ 0.14 | **0.79** ∓ 0.02 | 0.71 ∓ 0.01 | 0.63 ∓ 0.00 |
| | ACC | 0.58 ∓ 0.00 | 0.38 ∓ 0.09 | **0.81** ∓ 0.05 | 0.44 ∓ 0.03 | 0.34 ∓ 0.01 |
| UrbanGB | ARI | 0.63 ∓ 0.01 | 0.12 ∓ 0.05 | 0.09 ∓ 0.00 | **0.67** ∓ 0.05 | 0.49 ∓ 0.06 |
| | NMI | 0.68 ∓ 0.01 | 0.21 ∓ 0.07 | 0.24 ∓ 0.00 | 0.77 ∓ 0.01 | **0.81** ∓ 0.01 |
| | ACC | 0.45 ∓ 0.01 | 0.29 ∓ 0.02 | 0.29 ∓ 0.00 | 0.53 ∓ 0.02 | **0.54** ∓ 0.02 |

Table 2: La moyenne et l'écart type des trois métriques ARI, NMI et ACC, sur 10 exécutions. Le meilleur résultat dans chaque ligne est marqué en gras, et le deuxième meilleur est souligné.

## 4.3 Convergence

Maintenant, nous comparons les performances des modèles qui partagent les mêmes hypothèses: CGS et SubC sur les bases de données Synthetic 100K, Fashion-mnist et Balanced. La Figure 1 illustre la log-vraisemblance et le score ARI à chaque itération. Nous observons que notre algorithme converge presque à la même vitesse que l'algorithme centralisé CGS et est beaucoup plus rapide que SubC. Cela est dû au fait que notre algorithme est capable de découvrir de nouveaux *clusters* au niveau des *workers*. Or, dans SubC, le nombre de *clusters* est fixe à ce niveau, et les nouvelles composantes ne sont découvertes qu'au niveau du *master*. Ainsi, plus d'itérations sont nécessaires pour générer suffisamment de composantes qui représentent les données. Dans l'ensemble, notre algorithme converge rapidement et maintient un score ARI stable au fil des itérations. Alors que le CGS et SubC peuvent dégrader leur score comme on peut l'observer sur le jeu de données Fashion-Mnist.

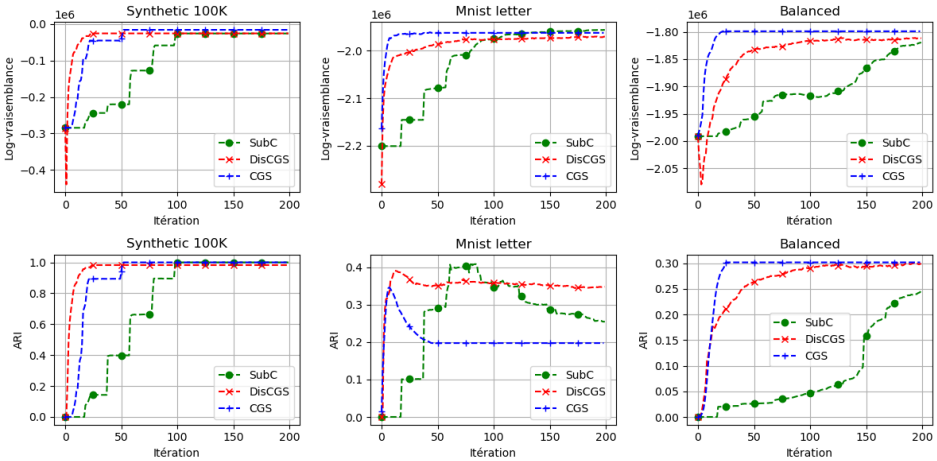



Figure 1: La log-vraisemblance et le score ARI à chaque itération.

## 4.4 Comparaison de la méthode distribuée et centralisée

Dans cette expérience, nous comparons le temps d'exécution et les performances de *clustering* de l'algorithme distribué (DisCGS) et centralisé (CGS). Nous exécutons les deux algorithmes sur des jeux de données synthétiques de tailles différentes (de $n = 20$K à $n = 100$K) générés à partir de $K = 10$ composantes gaussiennes de dimension 2. Les résultats présentés dans le tableau 3 montrent le gain significatif du temps d'exécution de l'approche distribuée sans que celle-ci compromette les performances de *clustering*. Par exemple, le temps de calcul est divisé par 200 sur le jeu de données de taille 100k. De plus, notre méthode obtient des scores plus élevés dans 80% des cas.

| Nombre d'observations | ARI | | NMI | | ACC | | Temps d'exécution | |
|---|---|---|---|---|---|---|---|---|
| | Dis. | Cen. | Dis. | Cen. | Dis. | Cen. | Dis. | Cen. |
| 20K | **0.99** | 0.89 | **0.99** | 0.96 | **0.99** | 0.89 | **66.53** | 1704.21 |
| 40K | 0.96 | **0.99** | 0.97 | **0.99** | **0.99** | 0.97 | **99.45** | 10898.28 |
| 60K | **0.91** | 0.89 | 0.92 | **0.96** | **0.92** | 0.89 | **135.76** | 25738.46 |
| 80K | **0.94** | 0.89 | **0.96** | **0.96** | **0.91** | 0.89 | **201.59** | 27492.08 |
| 100K | **0.91** | 0.89 | **0.94** | 0.89 | **0.91** | 0.89 | **207.58** | 44688.31 |

Table 3: Métriques de *clustering* et temps d'exécution obtenus par l'inférence distribuée (Dis.) et centralisée (Cen.) sur des jeux de données synthétiques de différentes tailles.

## 4.5 Passage à l'échelle

Dans cette expérience, nous utilisons $n = 10^6$ points de données générés à partir de $K = 10$ composantes gaussiennes bidimensionnelles. Nous exécutons notre inférence distribuée plusieurs fois en augmentant le nombre de cœurs de 8 à 48. La Figure 2 représente le temps d'exécution en fonction du nombre de cœurs. Nous observons que le temps d'exécution diminue significativement lorsque le nombre de cœurs augmente, montrant que notre algorithme passe efficacement à l'échelle lorsque le nombre de coeurs augmente.

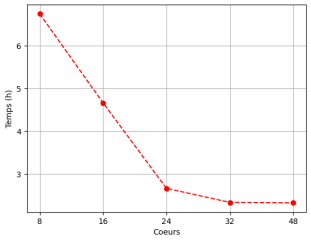



Figure 2: Temps d'exécution (h) en fonction du nombre de cœurs de DisCGS.

# 5 Conclusion et perspectives

Cet article introduit une nouvelle méthode MCMC distribuée pour les DPMM, spécialement conçue pour des données distribuées sur des machines indépendantes et hétérogènes, adaptée aux scénarios d'apprentissage horizontal fédéré. Les résultats expérimentaux sont prometteurs, démontrant une réduction significative du temps d'inférence tout en maintenant des performances satisfaisantes. Nos recherches en cours exploitent cette approche pour distribuerles modèles à blocs latents non paramétriques.

# 6 Remerciement

Projet soutenu par la Région Île-de-France dans le cadre du DIM AI4IDF. Je remercie Grid5000 pour avoir fourni les ressources de calculs nécessaires et la start-up HephIA pour l'échange sur les algorithmes distribués.

# Bibliographie

[1] J. Chang and J. W. Fisher III. Parallel sampling of dp mixture models using sub-cluster splits. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[2] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. Emnist: Extending mnist to handwritten letters. In *International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017.

[3] O. Dinari, A. Yu, O. Freifeld, and J. W. Fisher III. Distributed mcmc inference in dirichlet process mixture models using julia. In *19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 518–525, 2019.

[4] D. Dua and C. Graff. UCI machine learning repository.

[5] Y. Gal and Z. Ghahramani. Pitfalls in the use of parallel inference for the dirichlet process. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 208–216, Bejing, China, 22–24 Jun 2014. PMLR.

[6] H. Ge, Y. Chen, M. Wan, and Z. Ghahramani. Distributed inference for dirichlet process mixture models. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2276–2284, Lille, France, 07–09 Jul 2015.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. 1995.

[8] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[9] C. Jichan, K. Lee, and R. Kannan. Federated unsupervised clustering with generative models. *In AAAI 2022 International Workshop on Trustable, Verifiable and Auditable Federated Learning*, 2022.

[10] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence, 2016.

[11] H. H. Kumar, K. V R, and M. K. Nair. Federated k-means clustering: A novel edge ai based approach for privacy preservation. In *IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pages 52–56, 2020.

[12] D. Lovell, J. Malmaud, R. Adams, and V. Mansinghka. Clustercluster: Parallel markov chain monte carlo for dirichlet process mixtures. In *In Workshop on Big Learning, NIPS, 2012*.

[13] K. Meguelati, B. Fontez, N. Hilgert, and F. Masseglia. Dirichlet process mixture models made scalable and effective by means of massive distribution. pages 502–509, 04 2019.

[14] K. P. Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2$\sigma$2):16, 2007.

[15] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[16] K. Reda, L. Mustapha, A. Hanene, G. Etienne, and B. Djamel. Distributed collapsed gibbs sampler for dirichlet process mixture models in federated learning. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, 2024.

[17] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[18] S. Silvey. *Statistical Inference*. CRC Press, 2017.

[19] M. Stallmann and A. Wilbik. Towards federated clustering: A federated fuzzy c-means algorithm (FFCM). *CoRR*, abs/2201.07316, 2022.

[20] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 01 2002.

[21] R. Wang and D. Lin. Scalable estimation of dirichlet process mixture models on distributed data. In *International Joint Conference on Artificial Intelligence*, 2017.

[22] S. Williamson, A. Dubey, and E. Xing. Parallel Markov chain Monte Carlo for nonparametric mixture models. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 98–106, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[23] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[24] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*, 19(10):2761–2773, 2010.

# Intégration tardive de données multimodales par modèles à blocs stochastiques

Kylliann De Santiago [1] & Marie Szafranski [2] & Christophe Ambroise [3]

[1] *LaMME, France, kylliann.desantiago@univ-evry.fr*
[2] *ENSIIE, France, marie.szafranski@univ-evry.fr*
[3] *LaMME, France, christophe.ambroise@univ-evry.fr*

**Résumé.** Dans ce travail, nous présentons une méthode originale permettant d'agréger différentes sources d'information. Chaque partition est encodée par une matrice de co-appartenance des observations aux classes. Notre approche est fondée sur un mélange de modèles de blocs stochastiques multicouches pour conjointement définir des composantes de sources sur les matrices de co-appartenance similaires et partitionner les observations en différents groupes selon ces composantes. L'identifiabilité des paramètres du modèle est établie et un algorithme EM variationnel bayésien est proposé pour l'estimation de ces paramètres. Le cadre bayésien permet de sélectionner un nombre optimal de groupes et de composantes.

**Mots-clés.** Modèle à blocs stochastiques, apprentissage multivues, réseaux multicouches, cadre bayésien, vraisemblance complétée intégrée (ICL).

**Abstract.** In this work, we introduce an innovative method for aggregating multiple clusters originating from different sources of information. Each partition is represented by a co-membership matrix among observations. Our approach employs a mixture of multilayer Stochastic Block Models (SBM) to cluster co-membership matrices with similar information into components and to assign observations to distinct clusters, considering their specificities within the components. The identifiability of the model parameters is established, and a variational Bayesian EM algorithm is proposed for parameter estimation. The Bayesian framework facilitates the selection of an optimal number of clusters and components.

**Keywords.** Stochastic Block Model, Multiview clustering, Multilayer Network, Bayesian Framework, Integrated Classification Likelihood.

## 1 Introduction

La plupart des situations d'apprentissage font appel à différentes sources d'information, appelées ici *modalités* ou *vues*, telles que la vision, le toucher ou encore l'ouïe. Les objectifs de l'apprentissage machine multimodal ou multi-vues vont de l'apprentissage de nouvelles représentations, en passant par la traduction de textes ou encore la fusion d'information (cf. Zhao et al., 2017; Baltrušaitis et al., 2018; Cornuéjols et al., 2018, par exemple).

Les graphes constituent un moyen puissant et intuitif de représenter des systèmes complexes de relations entre individus. Ils fournissent une représentation efficace et informative

du système. La construction de graphes à partir de chaque vue permet d'utiliser l'apprentissage machine non supervisé dans une perspective de classification multimodale (Ektefaie et al., 2023).

Dans le cadre de la classification non supervisée, les algorithmes produisent souvent une partition ou une matrice d'appartenance $\mathbf{Z}$. Cette information, bien qu'utile, présente l'inconvénient de dépendre fortement du nombre de groupes choisi. Pour éviter ce problème, $\mathbf{Z}$ peut être transformé en une matrice d'adjacence $\mathbf{A}$ avec

$$A_{ij} = \begin{cases} 1, & \text{si les individus } i, j \text{ sont dans le même groupe,} \\ 0, & \text{sinon.} \end{cases}$$

Le terme de méta-classification(*meta clustering*) est utilisé pour désigner les approches permettant de combiner des partitions découvertes à l'aide de différentes méthodes. Parmi elles, la classification consensuelle (*consensus clustering*) est un algorithme de référence (Monti et al., 2003; Li et al., 2015; Liu et al., 2018). Lorsque ces approches sont fondées sur des modèles, elles permettent d'une part de construire un sous-typage final à partir des résultats déjà obtenus, mais aussi de mettre en lumière la redondance et / ou la complémentarité des sources d'information. En outre, utiliser un modèle permet aussi de disposer de critères d'évaluation de performance (log-vraisemblance, évidence, etc.) et, au moins dans le cadre bayésien, de critères de sélection de modèle (Biernacki et al., 2010).

Les modèles d'apprentissage varient en fonction de la stratégie de fusion de vues, précoce, intermédiaire ou tardive. Nous privilégierons ici la fusion tardive, où pour chaque vue, la matrice d'adjacence peut être obtenue par des algorithmes de classification efficaces et dédiés.

**Contribution.** À partir d'une classification indépendante de chaque vue, nous proposons d'apprendre une représentation coordonnée par le biais d'un modèle probabiliste. Notre modèle est un mélange de SBM multicouches associées à différentes sources d'information, avec une stratification des observations transversale comme illustré en Figure 1. L'algorithme associé est appelé mimi-SBM (De Santiago et al., 2024b, MIxture of Multilayer Integrator Stochastic Block Model). De plus, grâce au cadre bayésien, il est possible de développer un critère de sélection du modèle issu de la borne inférieure de l'évidence, à la fois pour le mélange de vues et le nombre de groupes. Enfin, l'identifiabilité des paramètres du modèle est établie et un algorithme EM bayésien variationnel est proposé pour estimer les paramètres.

# 2 Modèle de mélange de SBM multicouches

Notre modèle s'appuie sur un SBM avec deux ensembles de variables latentes correspondant respectivement à la structure des observations et à la structure des vues. Cette proposition se situe à la croisée du *Multilayer SBM* (MLSBM), qui recherche une matrice de co-appartenance traversante des observations sur toutes les couches, et du *Mixture of Multilayer SBM* (MMLSBM), qui recherche des motifs structurels au sein des couches.

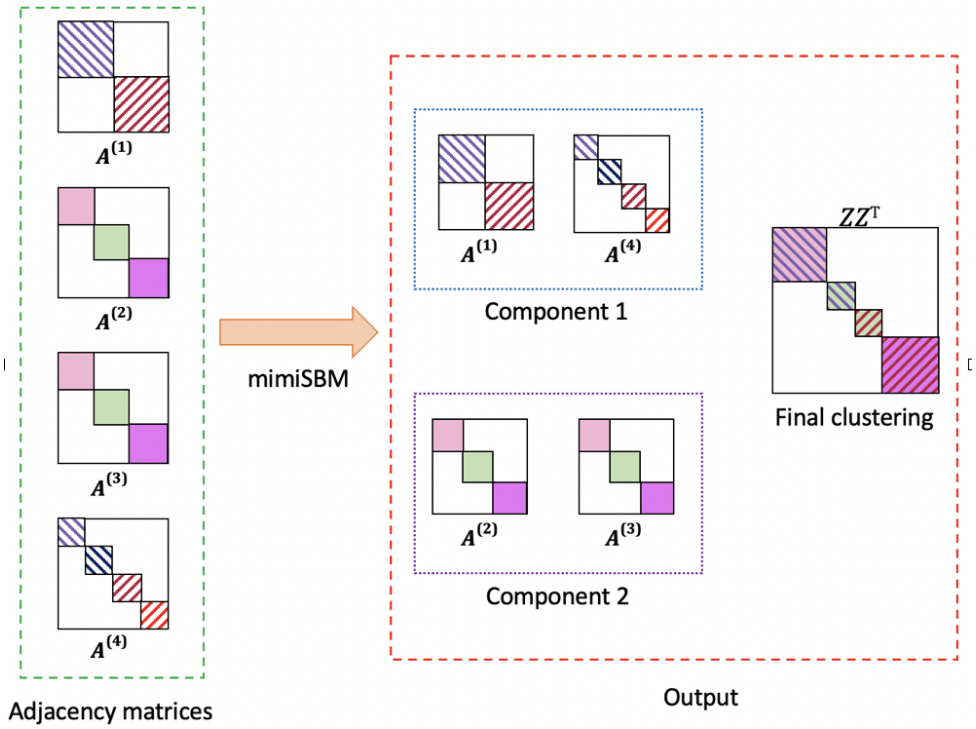



FIGURE 1 – Illustration de mimi-SBM. Gauche : 4 matrices d'adjacences $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(4)}$ provenant de quatre vues différentes organisées en deux composantes. Droite : identification des deux composantes à partir des vues (informations locales et complémentaires) et partition des observations décrites par la matrice d'appartenance $\mathbf{Z}$ (information globale et consensus).

**Observations.** Nous considérons, $\mathbf{A} \in \{0,1\}^{N \times N \times V}$, une représentation tensorielle des observations, où $N$ est le nombre d'observations (nœuds) et $V$ est le nombre de vues. Chacune des $V$ tranches de $\mathbf{A}$ est une matrice d'adjacence correspondant à un graphe $\mathcal{G}^v$. Le tenseur est donc un empilement de matrices d'adjacence relié aux graphes multi-vues $(\mathcal{G}^1, \cdots, \mathcal{G}^V)$ et nœuds correspondants. Soit $(i,j)$, une arête entre les observations $i$ et $j$. Par définition, $A_{ijv} = \mathbb{I}_{((i,j) \in E^v)}$, où $E^v$ est l'ensemble des arêtes du graphe $\mathcal{G}^v$.

**Structures latentes.** Soit $\mathbf{Z} \in \{0,1\}^{N \times K}$ la matrice d'appartenance des nœuds, où $K$ est le nombre de groupes traversant les vues. Pour le nœud $i$ et le groupe $k$, $Z_{ik} = \mathbb{I}_{(i \in k)}$. Notons aussi $\mathbf{W} \in \{0,1\}^{V \times Q}$, la matrice d'appartenance des vues, où $Q$ est le nombre de composantes du mélange des vues. Ainsi pour la vue $v$ et la composante $s$, $W_{vs} = \mathbb{I}_{(v \in s)}$.

## 2.1 Un mélange d'observations à travers un mélange de vues

Nous supposons que les $V$ vues sont générées par un modèle de mélange de $Q$ composantes, où chaque composante $s$ est un SBM. Nous supposons également que chaque ligne de la

matrice $\mathbf{W}$ suit une distribution multinomiale, $\mathbf{W}_v \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_Q))$, avec

$$\mathbb{P}(\mathbf{W} \mid \boldsymbol{\rho}) = \prod_{v=1}^{V} \prod_{s=1}^{Q} \rho_v^{W_{vs}}. \tag{1}$$

Nous supposons aussi une *structure traversante* à *toutes les vues* décrite par la variable latente $\mathbf{Z}$. En exploitant toutes les sources d'information disponibles, notre objectif est d'obtenir des composantes cohérentes sur l'ensemble des vues. On suppose ainsi que les individus proviennent d'un nombre $K$ de sous-populations. Chaque vecteur de classe latente pour l'observation $i$ suit une distribution multinomiale, avec $\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K))$, et

$$\mathbb{P}(\mathbf{Z} \mid \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{Z_{ik}}. \tag{2}$$

Enfin, chaque observation $A_{ijv}$ conditionnellement à la structure latente $\mathbf{Z}$ suit une distribution de Bernoulli : $A_{ijv} \mid Z_{ik} = 1, Z_{jl} = 1 W_{vs} = 1 \sim \mathcal{B}(\alpha_{kls})$. La probabilité de toutes les observations sachant les variables latentes $\mathbf{Z}$, $\mathbf{W}$, et un vecteur de paramètres $\boldsymbol{\Theta}$, est donc

$$\mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\Theta}) = \prod_{\substack{i=1,\\ i<j}}^{N} \prod_{\substack{k=1\\ l=1}}^{K} \prod_{v=1}^{V} \prod_{s=1}^{Q} \left( \alpha_{kls}^{A_{ijv}} \left(1 - \alpha_{kls}\right)^{1 - A_{ijv}} \right)^{Z_{ik} Z_{jl} W_{vs}}. \tag{3}$$

## 2.2 Identifiabilité

**Théorème 1** *Soient $N \geq \max(2K, 4Q)$ et $V \geq 2K$. Supposons que pour tout $k, l \in \{1, \ldots, K\}$ et chaque $s \in \{1, \ldots, Q\}$, les coordonnées de $\left(\boldsymbol{\pi}^T \boldsymbol{\alpha}_{k..} \boldsymbol{\rho}\right)$ sont toutes différentes, $(\boldsymbol{\pi}^T \boldsymbol{\alpha}_{..s} \boldsymbol{\pi})_{s=1:Q}$ sont distinctes, et chaque $(\boldsymbol{\alpha}_{kl.} \boldsymbol{\rho})_{k,l=1:K}$ est différent. Alors, les paramètres du mimi-SBM $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ sont identifiables.*

**Preuve 1** *La preuve de ce théorème est disponible dans (De Santiago et al., 2024a).*

## 2.3 Formulation bayesienne

Les modèles bayésiens offrent un cadre naturel pour incorporer des connaissances *a priori* permettant d'améliorer la précision de la structure estimée, en particulier lorsque les données disponibles sont limitées ou bruitées. Dans ce contexte, la définition des distributions conjuguées choisies à la fois pour la proportion du mélange et les proportions des blocs s'appuie sur les travaux de Latouche et al. (2012). Les lois *a priori* conjuguées conduisent à des distributions *a posteriori* explicites, où Dir(.) désigne la loi de Dirichlet :

$$\mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_K^0)) \;=\; \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta}^0), \tag{4}$$

$$\mathbb{P}(\boldsymbol{\rho} \mid \boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_Q^0)) \;=\; \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}^0), \tag{5}$$

$$\mathbb{P}(\boldsymbol{\alpha} \mid \boldsymbol{\eta}^0 = (\eta_{kls}^0), \boldsymbol{\xi}^0 = (\xi_{kls}^0)) \;=\; \prod_{k,k<l} \prod_{s} \mathrm{Beta}(\alpha_{kls}; \eta_{kls}^0, \xi_{kls}^0). \tag{6}$$

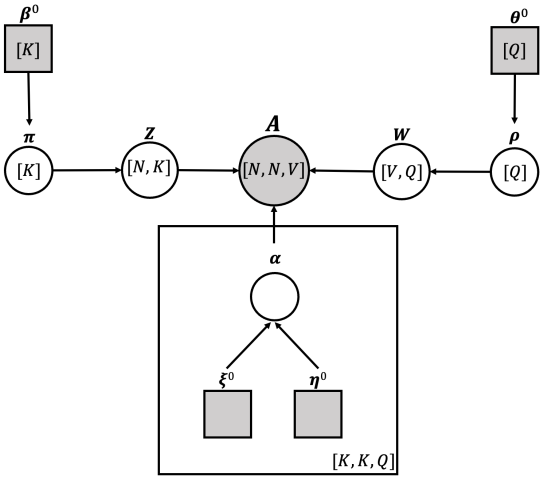



FIGURE 2 – Illustration du mimi-SBM avec les notations bayesiennes.

Les paramètres $\boldsymbol{\beta}^0, \boldsymbol{\theta}^0, \boldsymbol{\eta}^0, \boldsymbol{\xi}^0$ sont choisis selon les lois *a priori* de Jeffreys, qui sont souvent considérées comme non informatives ou faiblement informatives. Elles n'introduisent pas d'hypothèses *a priori* fortes ou de biais marqués dans l'analyse.

Pour la distribution de Dirichlet, fixer $\beta_k^0$ et $\theta_s^0$ à $1/2$ revient à considérer une distribution *a priori* de Jeffreys. De même, pour la distribution Beta, $\eta_{kls}^0$ et $\xi_{kls}^0$ peuvent être choisis comme étant égaux à $1/2$ pour tous les indices $k$, $l$, et $s$ correspondants.

# 3 Algorithme EM variationnel pour le mimi-SBM

La vraisemblance marginale dans les modèles de blocs stochastiques s'exprime par :

$$\mathbb{P}(\mathbf{A}) = \sum_{\mathbf{Z}} \sum_{\mathbf{W}} \int \int \int \mathbb{P}(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) \; d\boldsymbol{\alpha} \; d\boldsymbol{\pi} \; d\boldsymbol{\rho}. \tag{7}$$

Le calcul des intégrales de cette vraisemblance marginale ne présente pas de solution analytique. De plus, les sommes sur $\mathbf{Z}$ et $\mathbf{W}$ deviennent très difficiles à calculer lorsque le nombre de paramètres ou d'observations est conséquent. Pour contourner cela, on utilise généralement l'approximation de distributions postérieures complexes réalisée soit par échantillonnage (Monte-Carlo par chaînes de Markov ou des approches similaires), soit par l'inférence bayésienne variationnelle introduite par Attias (1999).

## 3.1 Borne inférieure de l'évidence

L'inférence variationnelle est efficace du point de vue computationnel et extensible pour les grands ensembles de données, et elle fonctionne particulièrement bien pour les modèles SBM. Le problème est formulé comme une tâche d'optimisation, où l'objectif est de trouver la meilleure approximation de la véritable distribution postérieure. Ce cadre d'optimisation

permet un calcul efficace des paramètres variationnels en maximisant une borne inférieure de la log-vraisemblance, connue sous le nom de borne inférieure de l'évidence (ELBO). Des techniques d'optimisation telles que la descente de gradient stochastique (SGD) ou l'algorithme d'EM peuvent être utilisées pour trouver les paramètres variationnels optimaux.

La distribution $\mathbb{P}(\mathbf{Z}, \mathbf{W}|\mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})$ étant incalculable dans le cadre des SBM, nous allons approcher l'ensemble de cette distribution. Soit une distribution variationelle $q$ sur $\{\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\}$, nous pouvons décomposer la log-vraisemblance marginale en deux parties : la borne inférieure de l'Évidence (ELBO) et la divergence de Kullback-Leibler **KL** entre la distribution variationnelle et la distribution postérieure :

$$\begin{aligned} \log P(\mathbf{A}) &= \mathbb{E}_q\left[\log \frac{P\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right)}{P(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}|\mathbf{A})}\right] && (8) \\ &= \underbrace{\mathbb{E}_q\left[\log \frac{P\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right)}{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})}\right]}_{ELBO=\mathcal{L}(q(.))} + \underbrace{\mathbb{E}_q\left[\log \frac{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})}{P(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho} \mid \mathbf{A})}\right]}_{\mathbf{KL}(q(\cdot)\|\,\mathbb{P}(\cdot|\mathbf{A}))} && (9) \end{aligned}$$

où $\mathbf{KL}\left(q(.) \mid \mathbb{P}(\cdot \mid \mathbf{A})\right) = -\mathbb{E}_q\left[\log \frac{p}{q}\right] \geq -\log \mathbb{E}_q\left[\frac{p}{q}\right] \geq 0$ selon l'inégalité de Jensen.

L'ELBO est donnée par

$$\mathcal{L}\left(q(\cdot)\right) = \sum_{\mathbf{Z},\mathbf{W}} \int\int\int q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) \log \frac{p\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right)}{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})}\, d\boldsymbol{\alpha}\, d\boldsymbol{\pi}\, d\boldsymbol{\rho}. \tag{10}$$

La distribution variationnelle est généralement choisie dans une famille de distributions plus facile à manipuler, comme la famille exponentielle. Les paramètres de la distribution variationnelle sont ensuite ajustés pour réduire la divergence de Kullback-Leibler par rapport à la distribution postérieure. Si $q(\cdot)$ est exactement égale à $p(\cdot|\mathbf{A})$, le terme **KL** est égal à 0, et l'ELBO est maximisée. Par une approximation du champ moyen, on définit $q(\cdot)$ comme :

$$\begin{aligned} q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) &= \prod_{i=1}^{N} q(\mathbf{Z}_i) \prod_{v=1}^{V} q(\mathbf{W}_v) \prod_{s=1}^{Q} \prod_{k,k\leq l}^{K} q(\alpha_{kls})\; q(\boldsymbol{\pi})\; q(\boldsymbol{\rho}), \\ &= \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta})\; \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}) \prod_{i=1}^{N} \mathcal{M}(\mathbf{Z}_i; 1, \boldsymbol{\tau}_i) \prod_{v=1}^{V} \mathcal{M}(\mathbf{W}_i; 1, \boldsymbol{\nu}_v) \qquad (11) \\ &\quad \prod_{s=1}^{Q} \prod_{k,k\leq l}^{K} \mathrm{Beta}(\alpha_{kls}; \eta_{kls}, \xi_{kls}), \end{aligned}$$

où $\tau_{ik}$ (resp. $\nu_{vs}$) sont les paramètres variationels indiquant la probabilité que l'individu $i$ (resp. une vue $v$) appartienne au groupe $k$ (resp. à la composante $s$).

D'après (10), avec $\Gamma(.)$ désignant la fonction gamma et étant donné la distribution $q(\cdot)$

choisie, l'ELBO est donné par

$$\begin{aligned}\mathcal{L}\left(q(\cdot)\right) = &\log\left\{\frac{\Gamma\left(\sum_{k=1}^{K}\beta_k^0\right)\prod_{k=1}^{K}\Gamma\left(\beta_k\right)}{\Gamma\left(\sum_{k=1}^{K}\beta_k\right)\prod_{k=1}^{K}\Gamma\left(\beta_k^0\right)}\right\} + \log\left\{\frac{\Gamma\left(\sum_{s=1}^{Q}\theta_s^0\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s\right)}{\Gamma\left(\sum_{s=1}^{Q}\theta_s\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s^0\right)}\right\} \\ &+ \sum_{k\leq l}^{K}\sum_{s=1}^{Q}\log\left\{\frac{\Gamma\left(\eta_{kls}^0+\xi_{kls}^0\right)\Gamma\left(\eta_{kls}\right)\Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls}+\xi_{kls}\right)\Gamma\left(\eta_{kls}^0\right)\Gamma\left(\xi_{kls}^0\right)}\right\} \\ &- \sum_{i}^{N}\sum_{k}^{K}\tau_{ik}\log\tau_{ik} \; - \sum_{v}^{V}\sum_{s}^{Q}\nu_{vs}\log\nu_{vs}.\end{aligned} \tag{12}$$

Cette expression, également appelée *Integrated Likelihood variational Bayes* (ILvb, Latouche et al., 2012)), peut être utilisée pour la sélection de modèle.

### 3.2 Optimisation des paramètres de la borne inférieure

Nous utilisons un algorithme EM variationnel bayésien pour estimer les paramètres (cf. Algorithme 1). L'algorithme débute par l'initialisation des paramètres du modèle, puis effectue de manière itérative deux étapes : l'étape d'Espérance Variationnelle Bayésienne (étape VBE) et l'étape de Maximisation (étape M).

Dans l'étape VBE, les distributions variationnelles $q(\mathbf{Z}_i)$ et $q(\mathbf{W}_v)$ sont optimisées sur les variables latentes $\forall i \in \{1,\ldots,N\}$ et $\forall v \in \{1,\ldots,V\}$ afin d'approximer la vraie distribution postérieure. Dans l'étape M, les paramètres du modèle $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ et $\boldsymbol{\xi}$ sont mis à jour pour maximiser une borne inférieure sur la log-vraisemblance, sachant les paramètres calculés lors de l'étape VBE.

Il existe plusieurs techniques pour initialiser l'algorithme EM. Une approche prévalente consiste à initialiser aléatoirement les paramètres selon une distribution choisie. Néanmoins, cette méthode peut manquer de fiabilité et ne pas fournir de valeurs de départ satisfaisantes pour l'algorithme. Dans Stanley et al. (2016), les paramètres $(\tau_{ik})$ et $(\nu_{vs})$ sont initialisés avec les résultats d'un modèle de blocs stochastiques appliqué séparément sur chaque vue. Pour initialiser mimi-SBM, nous combinons, avec l'algorithme des K-means, les résultats de SBM appliqués indépendamment sur les $V$ vues.

## 4 Sélection de modèle

Dans le contexte de la classification, la sélection de modèle fait souvent référence au processus de détermination du nombre idéal de groupes pour un ensemble de données donné. Dans notre situation, la décision clé réside dans la sélection des valeurs appropriées pour $K$ et $Q$ afin de trouver un équilibre entre les performances et la complexité du modèle. Pour cela, plusieurs critères basés sur la log-vraisemblance pénalisée peuvent être utilisés, tels que le Critère d'Information d'Akaike (AIC, Akaike, 1998), le Critère d'Information Bayésien (BIC, Schwarz, 1978) et plus récemment le critère de Vraisemblance Complétée Intégrée (ICL,

**Algorithme 1** mimi-SBM

**Require:** Tenseur $\mathbf{A}$, Nombre de groupes $K$, Nombre de composantes $Q$, précision *eps*.
Initialisation : $\tau_{ik}^{(old)}$ et $\nu_{vs}^{(old)}$
**while** $\|\mathcal{L}\left(q^{new}(.)\right) - \mathcal{L}\left(q^{old}(.)\right)\| < eps$ **do**
**VBE-step**
Calculer $\tau_{ik}^{(new)}$ $\forall i \in \{1, \ldots, N\}$ et $\forall k \in \{1, \ldots, K\}$
Calculer $\nu_{vs}^{(new)}$ $\forall v \in \{1, \ldots, V\}$ et $\forall s \in \{1, \ldots, Q\}$
**M-step**
Optimiser $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ sachant les paramètres $(\tau_{ik}^{(new)})$ et $(\nu_{vq}^{(new)})$
**ELBO**
Calculer $\mathcal{L}\left(q^{new}(.)\right)$
**end while**

Biernacki et al., 2000). Nous considérons spécifiquement le critère ICL et ses pénalisations associées bien adapté aux modèles de mélange (Biernacki et al., 2010).

L'ICL est fondé sur la log-vraisemblance intégrée des paramètres sur les données complètes. De plus, si nous supposons l'indépendance des paramètres de la probabilité de connexion des composantes-groupes $\boldsymbol{\alpha}$, des paramètres du mélange de communautés $\boldsymbol{\pi}$ et des paramètres du mélange de vues $\boldsymbol{\rho}$, alors

$$\begin{aligned} \mathrm{ICL}(\mathbf{A}, K, Q) &= \log \mathbb{P}(\mathbf{A}, \mathbf{Z}, \mathbf{W} \mid K, Q), \\ &= \log \int_{\boldsymbol{\alpha}} \mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha})\, \mathbb{P}(\boldsymbol{\alpha}) d\boldsymbol{\alpha} + \log \int_{\boldsymbol{\pi}} \mathbb{P}(\mathbf{Z} \mid \boldsymbol{\pi})\, \mathbb{P}(\boldsymbol{\pi}) d\boldsymbol{\pi}, \\ &\quad + \log \int_{\boldsymbol{\rho}} \mathbb{P}(\mathbf{W} \mid \boldsymbol{\rho})\, \mathbb{P}(\boldsymbol{\rho}) d\boldsymbol{\rho}. \end{aligned} \tag{13}$$

Dans le cadre variationnel, $\mathbf{Z}$ et $\mathbf{W}$ doivent être estimés. Ainsi, $\hat{\mathbf{Z}}$ (resp. $\hat{\mathbf{W}}$) peut être choisi directement comme le vecteur des paramètres variationnels $\boldsymbol{\tau}$ (resp. $\boldsymbol{\nu}$) ou par un Maximum a Posteriori (MAP) :

$$\hat{\mathbf{Z}}_i = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\ \tau_{ik}.$$

En utilisant des approximations telles que la formule d'approximation de Stirling sur $\mathbb{P}(\boldsymbol{\pi})$ et $\mathbb{P}(\boldsymbol{\rho})$ et l'approximation asymptotique de Laplace sur $\mathbb{P}(\boldsymbol{\alpha})$, nous pouvons définir un *ICL approximatif* :

$$\begin{aligned} \mathrm{ICL}(\mathbf{A}, K, Q) &\approx \log\left(\mathbb{P}(\mathbf{A}, \hat{\mathbf{Z}}, \hat{\mathbf{W}} \mid K, Q)\right) - \mathrm{pen}(K, Q), \\ &\approx \mathcal{L}\left(q(.)\right) - \mathrm{pen}(K, Q). \end{aligned} \tag{14}$$

où $\mathrm{pen}(K, Q) = \dfrac{1}{2}\dfrac{K(K+1)}{2} Q \log\left(V \dfrac{N(N-1)}{2}\right) + \dfrac{1}{2}(K-1)\log(N) + \dfrac{1}{2}(Q-1)\log(V)$.

Rappelons d'abord que notre modèle est fondé sur des matrices d'adjacence non orientées (symétriques) où seules les matrices triangulaires supérieures sans la diagonale sont prise

en compte. L'*ICL approximatif* (14) est composé d'une partie dépendant du nombre de paramètres de $\boldsymbol{\alpha}$ et du nombre d'arêtes des graphes associés aux matrices d'adjacence et d'une autre partie fonction du nombre de degrés de liberté dans les paramètres des mélanges ($\boldsymbol{\pi}$, $\boldsymbol{\rho}$) et du nombre de variables liées.

Dans le cadre bayésien, avec des lois *a priori* conjuguées, il est possible de définir un ICL exact (Côme and Latouche, 2015). Il peut être obtenu à partir de l'ILvb (12) lorsque l'entropie des variables latentes est nulle et que l'algorithme EM est un *Classification EM* (CEM, Celeux and Govaert, 1992). En d'autres termes, les paramètres variationnels sont égaux à 1 s'ils sont le MAP et 0 sinon. Ainsi, cet *ICL exact* peut être défini comme :

$$\begin{aligned}\mathrm{ICL}_{\mathrm{exact}}(\mathbf{A}, K, Q) = \log\left\{\frac{\Gamma\left(\sum_{k=1}^{K}\beta_k^0\right)\prod_{k=1}^{K}\Gamma\left(\beta_k\right)}{\Gamma\left(\sum_{k=1}^{K}\beta_k\right)\prod_{k=1}^{K}\Gamma\left(\beta_k^0\right)}\right\} + \log\left\{\frac{\Gamma\left(\sum_{s=1}^{Q}\theta_s^0\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s\right)}{\Gamma\left(\sum_{s=1}^{Q}\theta_s\right)\prod_{s=1}^{Q}\Gamma\left(\theta_s^0\right)}\right\} \\ + \sum_{k\leq l}^{K}\sum_{s=1}^{Q}\log\left\{\frac{\Gamma\left(\eta_{kls}^0+\xi_{kls}^0\right)\Gamma\left(\eta_{kls}\right)\Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls}+\xi_{kls}\right)\Gamma\left(\eta_{kls}^0\right)\Gamma\left(\xi_{kls}^0\right)}\right\}. \qquad (15)\end{aligned}$$

Par ailleurs, il est possible d'utiliser directement les paramètres variationnels, à la place du CEM, et de dériver ainsi un *ICL variationnel* à partir du critère précédent.

# Conclusion

Cet article décrit une nouvelle méthode de modèle de mélange de SBM multicouches et son algorithme associé *mimi-SBM*. Afin d'obtenir une borne inférieure de l'évidence calculable, une approche variationelle bayésienne a été utilisée, où chaque paramètre du modèle est estimé à l'aide d'un algorithme EM bayésien variationnel. De plus, le cadre bayésien permet de développer une stratégie de sélection de modèle. Enfin, une preuve de l'identifiabilité des paramètres est établie dans la pré-publication soumise de De Santiago et al. (2024a).

# Références

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Hagai Attias. A variational baysian framework for graphical models. *Advances in neural information processing systems*, 12, 1999.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning : A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2) : 423–443, 2018.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7) :719–725, 2000.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Exact and monte carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11) :2991–3002, 2010.

Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3) :315–332, 1992.

Etienne Côme and Pierre Latouche. Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15(6) :564–589, 2015.

Antoine Cornuéjols, Cédric Wemmert, Pierre Gançarski, and Younès Bennani. Collaborative clustering : Why, when, what and how. *Information Fusion*, 39 :81–95, 2018.

Kylliann De Santiago, Marie Szafranski, and Christophe Ambroise. Mixture of multilayer stochastic block models for multiview clustering. *arXiv preprint arXiv :2401.04682*, 2024a.

Kylliann De Santiago, Marie Szafranski, and Christophe Ambroise. mimisbm : Mixture of multilayer integrator stochastic block models. CRAN, Package R, 2024b. https ://cran.r-project.org/package=mimiSBM.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4) :340–350, 2023.

P. Latouche, É. Birmele, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1) :93–115, 2012.

Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Liangchen Liu, Feiping Nie, Arnold Wiliem, Zhihui Li, Teng Zhang, and Brian C Lovell. Multi-modal joint clustering with application for unsupervised attribute discovery. *IEEE Transactions on Image Processing*, 27(9) :4345–4356, 2018.

S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering : a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52 :91–118, 2003.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Natalie Stanley, Saray Shai, Dane Taylor, and Peter J Mucha. Clustering network layers with the strata multilayer stochastic block model. *IEEE transactions on network science and engineering*, 3(2) :95–105, 2016.

J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview : Recent progress and new challenges. *Information Fusion*, 38 :43–54, 2017.

# A Bayes Factor Approach for Gene-Based Analysis of Rare Variants Combining Conjugate Priors and Bayesian Variable Selection

Laurent Briollais[1,2]

[1] *Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada, laurent@lunenfeld.ca*

[2] *Biostatistics division, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*

**Abstract.** A common approach for detecting rare variants (RVs) associated with complex human diseases is to perform a gene-based or a region-based test of association. However, including all the RVs within a gene-based test might reduce its power since most RVs are not associated with the outcome of interest. As a shift to this paradigm, we propose to add a variable selection step to choose the RVs that compose the gene-based test statistic as a way to enhance the power of the test. We propose a Bayes Factor (BF) test statistic derived from the generalized linear model and its conjugate prior where functional annotation at the RV level can easily be integrated and the extent in prior belief can also be accommodated. A key component of our approach is the selection of the important RVs within a gene, which is performed through a novel scalable birth-death MCMC algorithm. Through simulation studies, we show that the proposed BF outperformed competing approaches both in terms of gene ranking and power to detect gene-based associations. The power of BF was improved by the use of functional annotation but interestingly, even when no annotation was included, substantial power gain was obtained from the variable selection procedure. Our application to a large whole-exome sequencing data set comparing 1,658 individuals with lung cancer to 1,492 healthy controls was able to identify new genes associated with lung cancer and pointed towards interesting cancer-related pathways.

**Keywords.** Bayes Factor; Generalized linear model; Bayesian variable selection; Gene-based analysis; Rare variant; Sequencing Studies

## 1 Introduction

With the increasing use of Next Generation Sequencing (NGS) technology in the past decade, several statistical methods for rare variant (RV) association testing have emerged. A first step is often to perform a gene-based (or region-based) test to narrow down potential genes/regions harbouring causal variants since an exhaustive search of single RV associations might lack power once correction for the millions of tests conducted is applied (Xu et al. (2021), Xu et al. (2023)). Burden and variance component (e.g. SKAT) test statistics have been the most popular gene-based methods. A strategy to increase the power of gene-based tests is

to incorporate some prior information at the RV level instead of at the gene level. Weighted approaches that use various annotation strategies to define the weights have the risk of prioritizing RVs not associated with the outcome while missing potentially causal RVs (i.e., when the weights do not correlate with the true association). As a shift to this paradigm, we propose here an alternative solution that is to perform a variable selection of RVs that should compose the gene-based test.

# 2 Model

## 2.1 Model setting

Our framework is based on a Bayesian generalized linear regression model and its conjugate prior proposed by Chen and Ibrahim (2023). For the individual $i$, $i \in \{1, .., n\}$, let $Y_i$ denote a phenotype (e.g., disease outcome) following an exponential family distribution $p(\theta_i, \tau)$, where $\theta_i$ denotes the canonical parameter and $\tau$ denotes the scale parameter. The density function of $Y_i$ is written as

$$p(y_i|\theta_i, \tau) = \exp\{a_i^{-1}(\tau)(y_i\theta_i - b(\theta_i)) + c(y_i, \tau)\}, \tag{1}$$

where $a_i$, $b$ and $c$ are known functions and determine a particular distribution type. For the simplicity purpose, we set $\tau = 1$ and $a_i^{-1}(\tau) = 1$ in equation (1), which leads to a form of natural exponential family distribution (e.g., Normal, Poisson, Gamma with known shape parameter, binomial and negative binomial distribution). Thus, the $Y_i$ density function in equation (1) can be simplified as

$$p(y_i|\theta_i) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i)\}. \tag{2}$$

A generalized linear model (GLM) with a $\theta$-link function $\theta(.)$ is constructed to assess the association between $k$ RVs ($G_{i1}, ..., G_{ik}$ denote $k$ genotypes for individual $i$) within a gene (or a specific region on a chromosome) and the phenotype $Y_i$,

$$\theta_i = \theta(\boldsymbol{X_i\beta}) = \theta(\beta_0 + \sum_{j=1}^{k} \beta_j G_{ij}). \tag{3}$$

In this model, $\boldsymbol{X} \equiv (\mathbb{1}, \boldsymbol{G}_1, ..., \boldsymbol{G}_k)$ is an $n \times (k+1)$ covariate matrix including a vector of ones $\mathbb{1}$ and $k$ RVs. The $i$th row of $\boldsymbol{X}$ is denoted as $\boldsymbol{X}_i \equiv (1, \boldsymbol{G}_{i.}) \equiv (1, G_{i1}, ..., G_{ik})$. The coefficients $(\beta_1, ..., \beta_k)$ represent the effect sizes of the $k$ RVs. The conjugate prior density function of $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, ..., \beta_k)^T$ is written as

$$\pi(\boldsymbol{\beta}|a_0, \boldsymbol{y_0}) \propto \exp\left[a_0\{\boldsymbol{y_0}'\theta(\boldsymbol{X\beta}) - \mathbb{1}'b(\theta(\boldsymbol{X\beta}))\}\right], \tag{4}$$

where $\boldsymbol{y_0} \equiv (y_{01}, ..., y_{0n})$ is an $n \times 1$ vector of prior parameters and $a_0 > 0$ is a scalar prior parameter. The $\boldsymbol{y_0}$ parameter could be interpreted as a prior guess for $E(\boldsymbol{Y})$, e.g. each individual's probability of disease when $Y_i$ is binary outcome. We define $y_{0i} = \exp(\boldsymbol{G}_{i.}\boldsymbol{w})/(1+\exp(\boldsymbol{G}_{i.}\boldsymbol{w}))$ for the binary phenotype $Y$, where $\boldsymbol{w} \equiv (w_1, ..., w_k)^T$ are weights given to the $k$

RVs. For instance, some genetic annotation priors can assign higher weights to coding variants (more likely to possess biological functions) than noncoding variants. The $a_0$ parameter is a precision parameter (or prior sample size) that quantifies the strength of our prior belief in $\boldsymbol{y_0}$ measuring individuals' prior prediction of $Y$.

Given this conjugate prior $(\boldsymbol{\beta}|a_0, \boldsymbol{y}_0) \sim D(\boldsymbol{y}_0, a_0)$ and equation (4), the posterior density of $\beta$ can be written as

$$(\boldsymbol{\beta}|\boldsymbol{y}, a_0, \boldsymbol{y}_0) \sim D(\frac{a_0\boldsymbol{y}_0 + \boldsymbol{y}}{a_0 + 1}, a_0 + 1),$$

and

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, a_0, \boldsymbol{y_0}) \propto \exp\left[(\boldsymbol{y} + a_0\boldsymbol{y_0})'\theta(\boldsymbol{X\beta}) - (a_0 + 1)\mathbb{1}'b(\theta(\boldsymbol{X\beta}))\right]. \tag{5}$$

In this study, we focus on a binary outcome, where $Y_i = 0$ or 1 represents healthy controls and patients with disease, respectively. Accordingly, assuming $Y_i \sim Bernoulli(p_i)$, a logistic regression is built for the RV association analysis,

$$\log(\frac{p_i}{1 - p_i}) = \beta_0 + \sum_{j=1}^{k} \beta_j G_{ij}. \tag{6}$$

## 2.2 Bayes factor for gene-based association

To assess the association between a group of RVs in the same gene and a disease outcome, we propose a test of hypothesis that compares a regression model including a chosen set of RVs to a model without any RV. We explain later how the set of RVs is chosen in the former model. Under the null hypothesis $H_0$, there is no association between any of the RVs and the outcome, so only the intercept $\beta_0$ is included in the regression model. Under the alternative hypothesis $H_1$, there is an association between a set of RVs and the outcome. Under $H_1$, assuming that a regression model $M$ is built, $G_{(M)}$ denotes the set of RVs and $k_M = |G_{(M)}|$ represents the total number of RVs in the model $M$. We write the covariates as $X_{(M)} = (\mathbb{1}, G_{(M)1}, ..., G_{(M)k_M})$ and vector of coefficients $\boldsymbol{\beta}_{(M)} \equiv (\beta_0, \boldsymbol{\beta}_{k_M})^T$, which include an intercept, $\beta_0$, and the coefficients associated with the RVs, $\boldsymbol{\beta}_{k_M} \equiv (\beta_{(M)1}, ..., \beta_{(M)k_M})^T$.

The Bayes factor (BF) plays a dual role in this framework. First, it is used to select the best model (i.e., the model that includes the best subset of RVs) by comparing alternative models using the scalable birth-death MCMC (SBDMCMC) algorithm, that we recently developed (Wang et al. (2023)). Second, once the best model has been identified, it is compared to a null model without any RV and the resulting BF is used as a gene-based association test statistic. Following Chen et al. (2008), the BF comparing the (best) model under $H_1$ (model $M$) to the model under $H_0$ (null model) can be defined as

$$BF_{M0} = \frac{C_{H_1}(D)/C_{H_0}(D)}{C_{H_1}(\boldsymbol{y}_0)/C_{H_0}(\boldsymbol{y}_0)}, \tag{7}$$

where $C_{H_1}(D)$ and $C_{H_1}(\boldsymbol{y}_0)$ represent normalizing constants of $\boldsymbol{\beta}$ posterior distribution $C_{H_1}(D) = \int \pi(\boldsymbol{\beta}_{(M)}|\boldsymbol{y}, \boldsymbol{y}_0, a_0)d\boldsymbol{\beta}_{(M)}$ and prior distribution $C_{H_1}(\boldsymbol{y}_0) = \int \pi(\boldsymbol{\beta}_{(M)}|\boldsymbol{y}_0, a_0)d\boldsymbol{\beta}_{(M)}$, under $H_1$,

respectively; while $C_{H_0}(D)$ and $C_{H_0}(\boldsymbol{y}_0)$ represent normalizing constants of $\boldsymbol{\beta}$ posterior distribution $C_{H_0}(D) = \int \pi(\beta_0|\boldsymbol{y}, \boldsymbol{y}_0, a_0)d\beta_0$ and prior distribution $C_{H_0}(\boldsymbol{y}_0) = \int \pi(\beta_0|\boldsymbol{y}_0, a_0)d\beta_0$, under $H_0$, respectively.

## 2.3 RV selection algorithm

A key component of our approach is the selection of the important RVs within a gene. Since the number of RVs within each gene is large, pairwise comparison between models is impossible. MCMC approaches are commonly used to find the model with the largest posterior probability. For this problem, we propose here to use a more efficient local Bayesian variable selection method that we recently developed (Wang et al. (2023)), the SBDMCMC algorithm .

The SBDMCMC is a continuous time Markov process in Bayesian model selection problems. As given in Section 2.2, we would like to use the SBDMCMC algorithm to select the best covariates (RVs) from the logistic regression model. In the logistic regression (6), the model space is denoted as the power set of $G$: $2^G$, and $G = (G_1, ..., G_k)$ is the full set of $k$ RVs. For any $M \in 2^G$, $G_{(M)}$ denotes the set of RVs in model $M$. The SBDMCMC process explores the model space by adding and removing variables (RVs) corresponding to birth and death jumps. Given the current model $M_1$ and its RV set $G_{(M_1)}$, the birth and death events are defined by the following independent Poisson processes:

- Birth event: each variable $G_r \notin G_{(M_1)}$ is born independently of other variables as a Poisson process with rate $B_r(M_1)$. If this birth event of variable $G_r$ happens, the process jumps to the new state $M_2$ with $G_{(M_2)} = G_{(M_1)} \cup G_r$.
- Death event: each variable $G_s \in G_{(M_1)}$ dies independently of other variables as a Poisson process with rate $D_s(M_1)$. If this death event of variable $G_s$ happens, the process jumps to the new state $M_2$ with $G_{(M_2)} = G_{(M_1)} \setminus G_s$.

The waiting time to the next birth/death jump from the current model $M_1$ follows an exponential distribution with mean

$$w(M_1) = \frac{1}{\sum_{G_r \notin G_{(M_1)}} B_r(M_1) + \sum_{G_s \in G_{(M_1)}} D_s(M_1)},$$

and the probability of the birth and death events are respectively

$$\begin{aligned} p_{M_1}(G_r) &= \frac{B_r(M_1)}{\sum_{G_{r'} \notin G_{(M_1)}} B_{r'}(M_1) + \sum_{G_s \in G_{(M_1)}} D_s(M_1)}, \quad G_r \notin G_{(M_1)}, \\ q_{M_1}(G_s) &= \frac{D_s(M_1)}{\sum_{G_r \notin G_{(M_1)}} B_r(M_1) + \sum_{G_{s'} \in G_{(M_1)}} D_{s'}(M_1)}, \quad G_s \in G_{(M_1)}. \end{aligned} \tag{8}$$

Assuming the posterior probability of model $M$ as $P(M|D) \equiv P(G_{(M)}|D)$, the birth and death rates are computed as

$$B_r(M_1) = \frac{1}{k}\frac{P(G_{(M_1)} \cup G_r|D)}{P(M_1|D)} = \frac{1}{k}\frac{P(M_2|D)}{P(M_1|D)}, \quad \forall G_r \notin G_{(M_1)}, G_{(M_2)} = G_{(M_1)} \cup G_r \tag{9}$$

and

$$D_s(M_1) \;=\; \tfrac{1}{k}\tfrac{P(G_{(M_1)}\setminus G_s|D)}{P(M_1|D)} = \tfrac{1}{k}\tfrac{P(M_2'|D)}{P(M_1|D)}, \quad \forall G_s \in G_{(M_1)}, G_{(M_2')} = G_{(M_1)} \setminus G_s, \tag{10}$$

respectively, which are based on the ratio between the posterior probability of the new model ($M_2$) and old model ($M_1$), $\frac{P(M_2|D)}{P(M_1|D)} = \frac{P(D|M_2)\pi(M_2)}{P(D|M_1)\pi(M_1)}$ (description of $M_2'$ as a new model is same as $M_2$, and it is omitted for simplification hereafter). Here, $\pi(M_1)$ and $\pi(M_2)$ denote the prior probability of model $M_1$ and $M_2$, respectively. Further, $BF_{21} \equiv \frac{P(D|M_2)}{P(D|M_1)}$ is the Bayes factor between model $M_2$ and $M_1$, which can be computed exactly using the conjugate prior as given in section 2.1 and 2.2. In order to avoid over-fitting, we can apply a sparse model space prior for our SBDMCMC algorithm by using $\pi(M) \propto \alpha^{k_M}, \quad \alpha \in (0, 1]$, where $k_M$ is the number of RVs in model $M$. The parameter $\alpha$ controls the sparsity of the prior distribution. The smaller the value $\alpha$ is, the larger probability a sparse model gets over a dense model. If $\alpha = 1$, all the models are equally distributed in the model space, which is the assumption we used in our simulations and read data analysis.

After applying the SBDMCMC algorithm to generate a sequence of samples from the model space: $M_1, M_2, \cdots, M_P$, we calculate their corresponding posterior probability $P(M_p|D)$, which is proportional to the waiting time $w(M_p)$. Thus, by applying the Bayesian model averaging, the posterior probability for RV $G_r$ being selected in the logistic regression is computed as

$$p(G_r) = \sum_{p=1}^{P} \mathbf{1}_{G_r \in G_{(M_p)}} P(M_p|D), \quad r = 1, ..., k.$$

This probability is also called the posterior inclusion probability(PIP). In this paper, we use 0.5 as the cutting value to decide the inclusion of each variable, i.e. RV $G_r$ is selected if $PIP(G_r) \geq 0.5$. One can use different cutting value to control the sparsity of variable selection results.

### 2.4 BF Computation and SBDMCMC algorithm

As described in section 2.3, we incorporated BF value in the SBDMCMC algorithm as the criterion for model fit to help select important RVs for a gene-based test, which we call as "BF-SBDMCMC" algorithm. Hereafter, BF is used as a short term for "BF-SBDMCMC" algorithm.

The BF-SBDMCMC algorithm is summarized as follow (see Algorithm 1 below), which is conducted for the RV selection purpose.

## 3 Simulation study

Our simulation studies show that the proposed BF-SBDMCMC approach outperforms competing approaches both in terms of gene ranking and power to detect gene-based associations.

The power of BF was improved by the use of functional annotation but interestingly, even when no annotation was included, substantial power gain was obtained from the variable selection procedure.

# 4 Application

## 4.1 Design

We applied BF-SBDMCMC to the WES study of the International Lung Cancer Consortium (ILCCO ) data to identify new genes that are associated with lung cancer. Four independent substudies conducted at 4 sites are included in the ILCCO data, including Harvard University School of Public Health/Massachusetts General Hospital (426 cases vs. 270 controls), University Health Network and Mount Sinai Hospital in Toronto (259 cases vs. 258 controls), University of Liverpool in UK (64 cases vs. 69 controls) and International Agency for Research on Cancer (293 cases vs. 284 controls). Our application was able to identify new genes associated with lung cancer and pointed towards interesting cancer-related pathways.

# 5 Conclusion

Our BF approach adds to the current methodologies on RV gene-based (or region-based) association tests. It allows for an easy integration of functional annotations through the elegant formulation of the conjugate prior proposed for GLM and was developed here specifically for sequencing association studies. Besides, the extent of prior belief is parametrized in this formulation through $a_0$ and can be decided by the user based on the confidence in prior annotation. The power of the gene-based statistic is improved by the use of functional annotation(s) but interestingly, even when no annotation was included, substantial power gain was obtained from the variable selection procedure, which retains only RVs with the highest PIP to be considered in the BF test statistic. This is a key result since in sequencing studies, reliable information on RVs is not always available and many RVs have unknown biological significance. We will also discuss how our approach can be used to define historical priors in a replication study.

# Bibliographie

Chen, M-H. and Ibrahim, J.G. (2003), Conjugate priors for generalized linear models, *Statistica Sinica*, 13, pp. 461-476.

Chen, M-H. and Huang, L. and Ibrahim, J.G. and Kim, S. (2008), Bayesian variable selection and computation for generalized linear models with conjugate priors, *Bayesian analysis*, 3, pp. 585-614.

Wang, N. and Massam, H. and Gao, X. and Briollais, L. (2023), The scalable birth–death MCMC algorithm for mixed graphical model learning with application to genomic data integration, *Annals of Applied Statistics*, 17, pp. 1958-1983.

Xu, J. and Xu, W. and Briollais, L. (2021), A Bayes factor approach with informative prior for rare genetic variant analysis from next generation sequencing data, *Biometrics*, 77, pp. 316-328.

Xu, J. and Xu, W. and Choi, J. and Brhane, Y. and Christiani, D. et al. (2023), Large-scale whole exome sequencing studies identify two genes, CTSL and APOE, associated with lung cancer, *PLoS genetics*, 19, pp. e1010902.

**Algorithm 1** BF-SBDMCMC algorithm for RV selection.

**Data:** $\boldsymbol{Y}, \boldsymbol{G} = (G_1, G_2, \cdots, G_k)$

**Result:** Select the best covariates in $\boldsymbol{G}$ (RVs) to fit the data in the logistic regression (6)

**Step 1(Birth-death process part):** Given the current model $M_1$,

1. For each $G_r \notin G_{(M_1)}$, denote $G_{(M_2)} = G_{(M_1)} \cup G_r$. Compute the $BF_{21}$ as given in Section 2.3, and calculate the birth rate as

$$B_r(M_1) = \frac{1}{k} BF_{21} \frac{\pi(M_2)}{\pi(M_1)},$$

2. For each $G_s \in G_{(M_1)}$, denote $G_{(M_2')} = G_{(M_1)} \setminus G_j$. Compute the $BF_{21}$ as given in Section 2.3, and calculate the death rate as

$$D_s(M_1) = \frac{1}{k} BF_{21} \frac{\pi(M_2')}{\pi(M_1)},$$

3. Calculate the waiting time as

$$w(M_1) = \frac{1}{\sum_{G_r \notin G_{(M_1)}} B_r(M_1) + \sum_{G_s \in G_{(M_1)}} D_s(M_1)},$$

4. Simulate the birth/death jump based on the birth/death probabilities of equation (8), and jump to new model $M_2$.

**Step 2(Sampling part):** Starting from the empty model $M_0$, repeat Step 1 $P$ times or until the variable selection reaches the local mode, generating $P$ samples, $M_1$, $M_2$, ... $M_P$. In our simulations and application, we chose $P = 30$.

**Step 3(Variable selection part):**

1. Calculate the PIP for each variable in $G$:

$$p(G_r) = \sum_{p=1}^{P} \mathbf{1}_{G_r \in G_{(M_p)}} P(M_p|D), \quad r = 1, ..., k.$$

2. Select the random variable $G_r$ if $p(G_r) \geq 0.5$ .

# Modèle génératif hiérarchique pour la rentrée atmosphérique

Pierre Minvielle [1] & Audrey Giremus [2] & Vivien Loridan [3]

[1] *CESTA, DAM, CEA, France, pierre.minvielle@cea.fr*
[2] *IMS (Université de Bordeaux, CNRS, Bordeaux INP), France, audrey.giremus@u-bordeaux.fr*
[3] *CESTA, DAM, CEA, France, vivien.loridan@cea.fr*

**Résumé.** Lorsqu'une navette spatiale ou une capsule rentre dans les couches denses de l'atmosphère, elle décélère fortement tandis qu'une onde de choc se forme en amont du véhicule, provoquant une montée en température et d'importants transferts de chaleur à la paroi. Prévoir l'évolution aérodynamique du véhicule à la rentrée est alors un enjeu essentiel, par exemple pour prédire la zone atterrissage. Pour ce faire, il est possible de s'appuyer sur de la simulation numérique multi-physique ainsi que sur quelques expériences de rentrée atmosphérique. Elles renseignent quant à l'erreur de simulation de l'évolution aérodynamique et à sa variabilité. Pour prévoir une future rentrée atmosphérique, il faut recourir à un modèle génératif qui, à partir des quelques courbes observées, en produit de nouvelles. La difficulté vient du nombre faible de données. L'approche proposée s'inspire du modèle KOH (Kennedy O'Hagan), largement répandu en calibration bayésienne de code et reposant sur des processus gaussiens. Cela conduit à un modèle bayésien hiérarchique qui tient compte de la non-stationnarité de l'erreur de modèle. La distribution prédictive peut être efficacement échantillonnée au moyen d'un échantillonneur Monte Carlo Hamiltonien (HMC) dynamique, dénommé No-U-Turn Sampler (NUTS), amplement exploité en statistiques computationnelles. Adapté à des problèmes de dimension conséquente, il va bénéficier des expressions analytiques de la densité cible et de son gradient pour explorer efficacement l'espace. Les échantillons générés, faiblement corrélés, peuvent être par la suite exploités pour prédire l'évolution aérodynamique et la zone d'atterrissage, via l'approximation de Monte Carlo d'intégrales multidimensionnelles en grande dimension, et en particulier, pour produire des intervalles de crédibilités.

**Mots-clés.** apprentissage statistique, calibration bayésienne, processus gaussien, Monte Carlo Hamiltonien

**Abstract.** When a space shuttle or capsule re-enters the dense layers of the atmosphere, it decelerates sharply while a shock wave is formed upstream the vehicle, causing a rise in temperature and significant heat transfer to the wall. Predicting the aerodynamic evolution of the vehicle on re-entry is therefore an essential challenge, for example in determining the landing zone. To do this, it is possible to rely on multi-physics numerical simulation and on atmospheric re-entry experiments. They provide information on the simulation error of aerodynamic evolution and its variability. To predict future atmospheric re-entry, we need to use a generative model which, based on the few curves observed, generates new ones. The difficulty lies in the small amount of data available. The proposed approach is inspired by

the KOH (Kennedy O'Hagan) model, widely used in Bayesian code calibration and based on Gaussian processes. This leads to a hierarchical Bayesian model that takes account of the non-stationarity of the model error. The predictive distribution can be efficiently sampled by a dynamic Monte Carlo Hamiltonian (HMC) sampler, the so-called No-U-Turn Sampler (NUTS), widely exploited in computational statistics. Adapted to high-dimensional problems, it will benefit from analytical expressions of the target density and its gradient to efficiently explore space. The weakly correlated generated samples can be used afrerwards to predict aerodynamic evolution and the landing zone, via Monte Carlo approximation of high-dimensional multidimensional integrals, and in particular to produce credibility intervals.

**Keywords.** statistical learning, Bayesian calibration, Gaussian process, Hamiltonian Monte Carlo

# 1 Introduction

Un aéronef effectuant une rentrée atmosphérique évolue typiquement dans des régimes hypersoniques. Sous de telles conditions, une onde de choc apparaît autour du véhicule, au sein de laquelle l'air incident est fortement compressé et ralenti. Il en résulte une augmentation drastique de la température, qui génère d'importants transferts de chaleur entre la surface de l'objet et l'écoulement. Les flux thermiques intenses auxquels sont soumis les matériaux de paroi peuvent potentiellement mettre en péril l'intégrité du véhicule. Afin de prévenir ce type de dommages, le système de protection thermique est spécifiquement conçu pour absorber une partie importante de l'énergie incidente via sa dégradation graduelle durant la phase de rentrée atmosphérique. Ce phénomène d'ablation modifie la géométrie du bouclier thermique au cours de la rentrée, ce qui a des conséquences sur les propriétés aérodynamiques du module et, par conséquent, sur sa trajectoire globale. La prédiction de la zone probable d'atterrissage repose ainsi sur deux aspects. D'une part, il s'agit de restituer le plus fidèlement possible les phénomènes multi-physiques impliqués lors de la rentrée atmosphérique (écoulements hypersoniques, turbulence, ablation, évolution thermique au sein des matériaux etc...) via le développement et l'utilisation de codes dédiés aux calculs aérothermiques. D'autre part, un couplage à la mécanique du vol est nécessaire afin de mettre à jour la trajectoire conformément aux changements aérodynamiques subis par le véhicule.

En complément, la prédiction peut s'appuyer sur des expériences représentatives de rentrée atmosphérique. Elles permettent de recaler les modèles et informent quant à la variabilité aérodynamique. Ci-après, on se place dans un formalisme proche de Kennedy et O'Hagan (2001), répandu en calibration bayésienne de codes. Ce travail présente la prédiction de l'évolution aérodynamique par un modèle génératif hiérarchique. La première partie expose le problème de calibration de code. Les deux parties suivantes décrivent le modèle bayésien hiérarchique et l'approche numérique d'échantillonnage. Enfin, des résultats illustratifs sont fournis.

## 1.1 Calibration de code

Lors d'une rentrée atmosphérique, une des principales caractéristiques aérodynamiques est le coefficient balistique $\beta = S \cdot C_A/m$, avec $S$ la surface de référence, $m$ la masse du véhicule et $C_A$ le coefficient aérodynamique. Intervenant dans la force de traînée, $\beta$ détermine la décélération de l'objet lors de la rentrée atmosphérique. Le coefficient $\beta$ dépend du véhicule : il est fonction de l'entrée $\mathbf{x} \in \aleph$ dans l'espace des entrées (géométrie du véhicule, matériaux, etc.). Il est de plus fonction de l'altitude $h$. On peut donc le noter $\beta(\mathbf{x}, h)$. En faisant intervenir un code multi-physique déterministe $\mathcal{T}$, $\beta(\mathbf{x}, h)$ peut se décomposer de la manière suivante :

$$\beta(\mathbf{x}, h) = \mathcal{T}(\mathbf{x}, h) + \Delta\beta(\mathbf{x}, h) \tag{1}$$

où $\mathcal{T}(\mathbf{x}, h)$ est le coefficient balistique calculé pour l'entrée $\mathbf{x} \in \aleph$ et l'altitude $h$, et $\Delta\beta(\mathbf{x}, h)$ est l'erreur de modèle associée (inconnue). En transposant plus spécifiquement à chacune des expériences de rentrée atmosphérique, impliquant un véhicule différent décrit par l'entrée $\mathbf{x}_m \in \aleph$ $(m = 1 \cdots M)$, il en découle :

$$\beta(\mathbf{x}_m, h) = \mathcal{T}(\mathbf{x}_m, h) + \Delta\beta(\mathbf{x}_m, h) \tag{2}$$

A chaque expérience, on observe *indirectement* $\beta(\mathbf{x}_m, h)$, c'est-à-dire après estimation, reconstruction ou restitution, au moyen de capteurs proprioceptifs et extéroceptifs. On suppose que ces observations se font à certaines altitudes discrètes : $h_1, h_2, \cdots, h_K$. Soit $\mathbf{y}_m$ le vecteur des coefficients balistiques observés, il peut se décomposer de la manière suivante :

$$\mathbf{y}_m = \boldsymbol{\mathcal{T}}(\mathbf{x}_m) + \Delta\boldsymbol{\beta}_m + \boldsymbol{\varepsilon}_m \tag{3}$$

où $\Delta\boldsymbol{\beta}_m \triangleq [\Delta\beta(\mathbf{x}_m, h_1) \quad \Delta\beta(\mathbf{x}_m, h_2) \quad \cdots \quad \Delta\beta(\mathbf{x}_m, h_K)]^T$ est le vecteur des erreurs de modèle et $\boldsymbol{\varepsilon}_m \triangleq [\varepsilon_m(\mathbf{x}_m, h_1) \quad \varepsilon_m(\mathbf{x}_m, h_2) \quad \cdots \quad \varepsilon_m(\mathbf{x}_m, h_K)]^T$ est le vecteur des erreurs de restitution aux altitudes $h_1, h_2, \cdots, h_K$.

Enfin, on considère une future expérience de rentrée correspondant à l'entrée $\mathbf{x}_\infty \in \aleph$ :

$$\beta(\mathbf{x}_\infty, h) = \mathcal{T}(\mathbf{x}_\infty, h) + \Delta\beta(\mathbf{x}_\infty, h) \tag{4}$$

Comme l'expérience n'a pas encore eu lieu, on ne dispose pas encore d'observation associée. Le problème est le suivant : il s'agit de "calibrer" l'erreur de modèle $\Delta\beta(\mathbf{x}, h)$, c'est-à-dire déterminer quelles sont ses possibles valeurs, et prévoir l'erreur de modèle $\Delta\beta(\mathbf{x}_\infty, h)$, ou de manière équivalente $\beta(\mathbf{x}_\infty, h)$ au regard des observations $\mathbf{y}_{1:M} \triangleq \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_M\}$. Ces dernières sont illustrées sur la figure 1 pour $M = 5$ où l'on représente en particulier les courbes d'écarts relatifs entre la simulation et l'expérience.

## 1.2 Problème génératif à partir d'un faible nombre de données

Ce problème peut être reformulé sous la forme d'un problème génératif. Prévoir l'erreur de modèle $\Delta\beta(\mathbf{x}_\infty, h)$ revient à être capable de générer, à partir de ces quelques courbes,

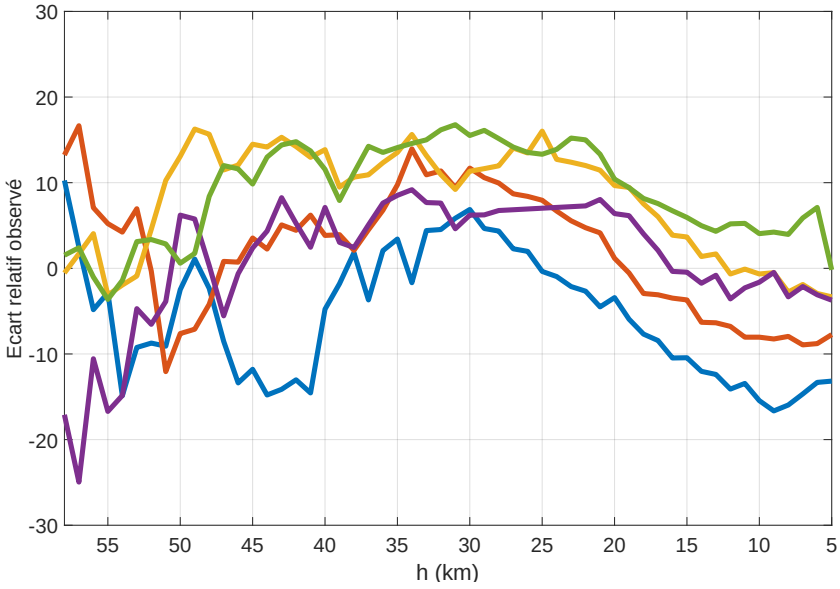



Figure 1: Ecarts relatifs observés du problème génératif ($M = 5$)

de nouvelles courbes que l'on souhaite représentatives. Autrement dit, d'un point de vue bayésien, on cherche à échantillonner aussi fidèlement que possible la densité *a posteriori* $p(\Delta\boldsymbol{\beta}_\infty|\mathbf{y}_{1:M})$. Ce problème génératif est difficile, en raison du faible nombre de données, et il ne peut être résolu au moyen des modèles génératifs à base de réseaux neuronaux (GAN, VAE, etc.). Sans information supplémentaire, on ne saurait dire par exemple si les seules $M$ courbes sont licites et issues d'une loi discrète, ou si elles sont des réalisations d'une loi continue. Cela nécessite donc de préciser la connaissance *a priori* et de faire le lien entre l'écart futur et les écarts passés $\Delta\boldsymbol{\beta}_m$. Tel est bien le rôle du modèle génératif hiérarchique, décrit ci-après.

# 2 Modèle génératif hiérarchique

## 2.1 Processus gaussien

Pareillement au modèle KOH de Kennedy O' Hagan (2001), on opte pour une modélisation par un processus aléatoire gaussien :

$$\Delta\beta(\mathbf{x}, h) \sim \mathcal{PG}(\mu(\mathbf{x}, h), k(\mathbf{x}, h, \mathbf{x}', h')) \tag{5}$$

où $\mu(\cdot,\cdot)$ est la fonction moyenne et $k(\cdot,\cdot,\cdot,\cdot)$ est le noyau.

**Stationnarité selon x** Le processus gaussien est supposé stationnaire au sens strict selon $\mathbf{x}$, c'est-à-dire que $\Delta\beta(\mathbf{x}, h)$ et $\Delta\beta(\mathbf{x}', h)$ ont même distribution de probabilité. A $\mathbf{x} = \mathbf{x}_m$ fixé,

l'erreur de modèle $\Delta\beta(\mathbf{x}_m, h)$ suit un processus gaussien indexé par $h \in \mathbb{R}$ ($m = 1, \cdots, M$) :

$$\Delta\beta(\mathbf{x}_m, h) \sim \mathcal{PG}(\mu(h), k(h, h')) \tag{6}$$

où $\mu(\cdot)$ est la fonction moyenne et $k(\cdot, \cdot)$ est le noyau ou fonction de covariance. Ainsi, le processus est stationnaire en fonction de l'expérience, c'est-à-dire que toutes les densités de probabilités sont supposées invariantes selon $m$. Formulé autrement, les véhicules ont des caractéristiques suffisamment proches pour que les écarts aient des distributions similaires. De même, on a naturellement :

$$\Delta\beta(\mathbf{x}_\infty, h) \sim \mathcal{PG}(\mu(h), k(h, h')) \tag{7}$$

**Non-stationnarité selon $h$** Au vu de la figure 1, on choisit de modéliser l'écart par un processus gaussien non-stationnaire en $h$, au premier et second ordre. Concernant la fonction de covariance, on adopte plus spécifiquement la généralisation non-stationnaire du noyau SE (Squared Exponential) :

$$k(h, h') = \sigma(h)\sigma(h')\sqrt{\frac{2\ell(h)\ell(h')}{\ell(h)^2 + \ell(h')^2}} exp\left(-\frac{(h - h')^2}{\ell(h)^2 + \ell(h')^2}\right) \tag{8}$$

où $\sigma(\cdot)$ et $\ell(\cdot)$ correspondent respectivement à l'écart-type et à la longueur de corrélation qui varient en fonction de l'altitude $h$. Un tel modèle fut tout d'abord introduit par Gibbs (1997), puis étendu ultérieurement par Paciorek et Schervish (2004).

On définit les vecteurs aléatoires $\Delta\boldsymbol{\beta}_m$ et $\Delta\boldsymbol{\beta}_\infty$, indépendants, tels que :

$$\Delta\boldsymbol{\beta}_m \triangleq [\Delta\beta(\mathbf{x}_m, h_1) \quad \Delta\beta(\mathbf{x}_m, h_2) \quad \cdots \quad \Delta\beta(\mathbf{x}_m, h_K)]^T \tag{9}$$

$$\Delta\boldsymbol{\beta}_\infty \triangleq [\Delta\beta(\mathbf{x}_\infty, h_1) \quad \Delta\beta(\mathbf{x}_\infty, h_2) \quad \cdots \quad \Delta\beta(\mathbf{x}_\infty, h_K)]^T \tag{10}$$

avec $K$ le nombre d'altitudes discrétisées. On constate alors que $\Delta\boldsymbol{\beta}_m \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$ et $\Delta\boldsymbol{\beta}_\infty \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$ où $\boldsymbol{\mu} \triangleq [\mu(h_1) \quad \mu(h_2) \quad \cdots \quad \mu(h_K)]^T$ est le vecteur moyenne et $\mathbf{P} \triangleq [k(h_i, h_j)]_{i=1\cdots K, j=1\cdots K}$ est la matrice de covariance dont les composantes sont :

$$\mathbf{P}_{ij} = \boldsymbol{\sigma}_i\boldsymbol{\sigma}_j\sqrt{\frac{2\boldsymbol{\ell}_i\boldsymbol{\ell}_j}{\boldsymbol{\ell}_i^2 + \boldsymbol{\ell}_j^2}} exp\left(-\frac{(h_i - h_j)^2}{\boldsymbol{\ell}_i^2 + \boldsymbol{\ell}_j^2}\right) \tag{11}$$

en notant $\boldsymbol{\sigma} \triangleq [\sigma(h_1) \quad \sigma(h_2) \quad \cdots \quad \sigma(h_K)]^T$ et $\boldsymbol{\ell} \triangleq [\ell(h_1) \quad \ell(h_2) \quad \cdots \quad \ell(h_K)]^T$.

## 2.2 Vraisemblance

Les erreurs de mesure, précédemment introduites dans l'eq. (3), sont modélisées par des vecteurs gaussiens : $\boldsymbol{\varepsilon}_m \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{R})$ où $\mathbf{R}$ est la matrice de covariance de mesure. On peut alors écrire :

$$\mathbf{y}_m | \Delta\boldsymbol{\beta}_m \sim \mathcal{N}(\boldsymbol{\mathcal{T}}(\mathbf{x}_m) + \Delta\boldsymbol{\beta}_m, \mathbf{R}) \tag{12}$$

## 2.3 Modèle bayésien hiérarchique

Pour garantir la positivité des hyperparamètres du modèle de processus gaussien, on introduit les vecteurs aléatoires de travail suivants :

$$\tilde{\boldsymbol{\sigma}} \triangleq [\mathrm{Log}\sigma(h_1) \quad \mathrm{Log}\sigma(h_2) \quad \cdots \quad \mathrm{Log}\sigma(h_K)]^T \tag{13}$$

$$\tilde{\boldsymbol{\ell}} \triangleq [\mathrm{Log}\ell(h_1) \quad \mathrm{Log}\ell(h_2) \quad \cdots \quad \mathrm{Log}\ell(h_K)]^T \tag{14}$$

Après regroupement, on aboutit à l'hyper-paramètre : $\boldsymbol{\theta} \triangleq [\boldsymbol{\mu}^T \quad \tilde{\boldsymbol{\sigma}}^T \quad \tilde{\boldsymbol{\ell}}^T]^T$ de dimension $3 \cdot K$. Il faut souligner que les vecteurs aléatoires $\Delta\boldsymbol{\beta}_m$ $(m = 1, \cdots, M)$ et $\Delta\boldsymbol{\beta}_\infty$ sont liés entre eux par l'intermédiaire des paramètres $\boldsymbol{\theta}$ du processus gaussien. Cela permet d'apprendre la loi de $\Delta\boldsymbol{\beta}_\infty$ à partir des $\Delta\boldsymbol{\beta}_m$ et donc des données $\mathbf{y}_m$ $(m = 1, \cdots, M)$. Avec l'approche hiérarchique bayésienne, il est possible de modéliser le manque d'information sur les paramètres d'une distribution *a priori* en recourant au paradigme de Bayes, c'est-à-dire en spécifiant une autre distribution sur ces paramètres (loi *hyper a priori)* : $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, où $p(\boldsymbol{\theta})$ est la densité de probabilité de $\boldsymbol{\theta}$. On choisit les lois *a priori* indépendantes suivantes : $\boldsymbol{\mu} \sim \mathcal{N}(0, \mathbf{P}^{SE}_{\beta_\mu,\gamma_\mu})$, $\tilde{\boldsymbol{\sigma}} \sim \mathcal{N}(\alpha_\sigma \cdot \mathbf{1}_K, \mathbf{P}^{SE}_{\beta_\sigma,\gamma_\sigma})$ et $\tilde{\boldsymbol{\ell}} \sim \mathcal{N}(\alpha_\ell \cdot \mathbf{1}_K, \mathbf{P}^{SE}_{\beta_\ell,\gamma_\ell})$, avec $\mathbf{P}^{SE}_{\beta,\gamma} \triangleq [k^{SE}_{\beta,\gamma}(h_i, h_j)]_{i=1\cdots K, j=1\cdots K}$. Le noyau SE (Squared Exponential) $k^{SE}_{\beta,\gamma}$, aussi dénommé gaussien, s'exprime classiquement par :

$$k^{SE}_{\beta,\gamma}(h, h') = \beta^2 \exp\left(-\frac{(h-h')^2}{2\gamma^2}\right) \tag{15}$$

Les paramètres $\beta_\mu, \gamma_\mu, \alpha_\sigma, \beta_\sigma, \gamma_\sigma, \alpha_\ell, \beta_\ell$ et $\gamma_\ell$ sont fixés de telle manière à induire des lois faiblement informatives. Sur la figure 2, on présente quelques réalisations *a priori* de $\Delta\boldsymbol{\beta}_m$ $(m = 1, \cdots, M)$ ou de manière équivalente de $\Delta\boldsymbol{\beta}_\infty$, à partir du modèle hiérarchique, pour des valeurs fixées plus loin dans la partie 4. Cela illustre la connaissance, notamment la non-stationnarité, qui est injectée dans la loi *a priori* du processus aléatoire.

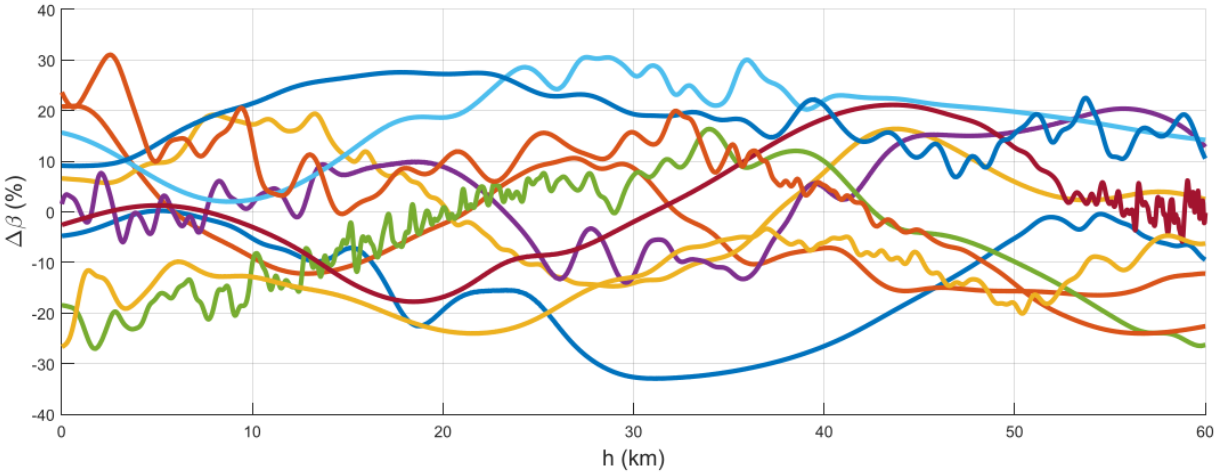



Figure 2: Réalisations du modèle hiérarchique

En complément, on choisit de simplifier drastiquement le modèle hiérarchique en faisant l'hypothèse d'indépendance entre erreurs de modèle $\Delta\boldsymbol{\beta}_m$ et $\Delta\boldsymbol{\beta}_{m'}$ $(m \neq m')$, conditionnellement à $\boldsymbol{\theta}$. De plus, de manière naturelle, on considère qu'il y a indépendance entre l' erreur de modèle $\Delta\boldsymbol{\beta}_m$ et l'erreur de mesure $\boldsymbol{\varepsilon}_m$.

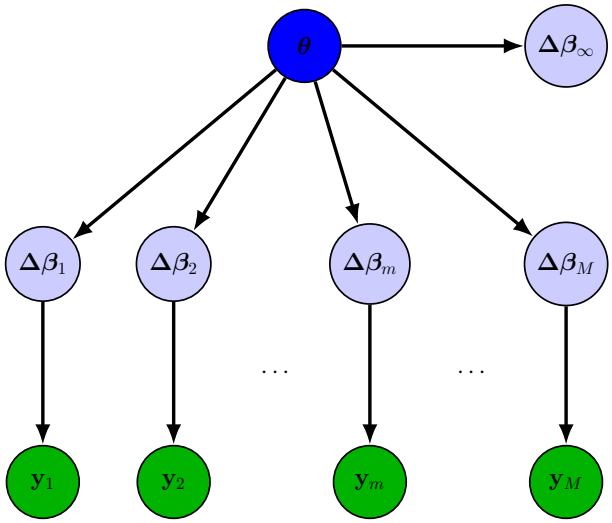



Figure 3: Modèle hiérarchique (propriété de Markov globale)

**Modèle graphique probabiliste** Le modèle graphique est constitué par le graphe orienté acyclique de la figure 3 qui représente les dépendances conditionnelles. On en déduit la décomposition suivante de la loi jointe :

$$p(\mathbf{y}_{1:M}, \Delta\boldsymbol{\beta}_{1:M}, \Delta\boldsymbol{\beta}_\infty, \boldsymbol{\theta}) = p(\mathbf{y}_{1:M}|\Delta\boldsymbol{\beta}_{1:M})p(\Delta\boldsymbol{\beta}_{1:M}|\boldsymbol{\theta})p(\Delta\boldsymbol{\beta}_\infty|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (16)$$

$$= \prod_{m=1}^{M} p(\mathbf{y}_m|\Delta\boldsymbol{\beta}_m)p(\Delta\boldsymbol{\beta}_m|\boldsymbol{\theta}) \cdot p(\Delta\boldsymbol{\beta}_\infty|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (17)$$

Plus spécifiquement, la loi conditionnelle $p(\mathbf{y}_m|\boldsymbol{\theta})$ s'écrit :

$$p(\mathbf{y}_m|\boldsymbol{\theta}) = \int p(\mathbf{y}_m|\Delta\boldsymbol{\beta}_m)p(\Delta\boldsymbol{\beta}_m|\boldsymbol{\theta})d\Delta\boldsymbol{\beta}_m \quad (18)$$

Par linéarité du modèle de mesure (12), on montre aisément que la loi est gaussienne :

$$p(\mathbf{y}_m|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}_m, \boldsymbol{\mathcal{T}}(\mathbf{x}_m) + \boldsymbol{\mu}, \mathbf{P} + \mathbf{R}) \quad (19)$$

# 3 Inférence bayésienne

S'appuyant sur le modèle hiérarchique proposé, on précise ci-après le problème inférentiel que l'on cherche à résoudre ainsi que l'approche numérique que l'on met en œuvre.

## 3.1 Distribution prédictive

Rappelons que le problème génératif requiert d'échantillonner $p(\Delta\boldsymbol{\beta}_\infty|\mathbf{y}_{1:M})$, la distribution prédictive. Après marginalisation, elle s'exprime par :

$$p(\Delta\boldsymbol{\beta}_\infty|\mathbf{y}_{1:M}) \;=\; \int p(\Delta\boldsymbol{\beta}_\infty|\boldsymbol{\theta},\mathbf{y}_{1:M})p(\boldsymbol{\theta}|\mathbf{y}_{1:M})d\boldsymbol{\theta} \tag{20}$$

On peut alors faire l'approximation suivante : $p(\Delta\boldsymbol{\beta}_\infty|\mathbf{y}_{1:M}) \approx \frac{1}{N}\sum_{n=1}^{N} p(\Delta\boldsymbol{\beta}_\infty|\boldsymbol{\theta}^{(n)},\mathbf{y}_{1:M})$, si l'on est capable de produire des échantillons $\boldsymbol{\theta}^{(n)}$ de la distribution *a posteriori* $p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$. Or, $p(\Delta\boldsymbol{\beta}_\infty|\boldsymbol{\theta}^{(n)},\mathbf{y}_{1:M}) = \mathcal{N}(\Delta\boldsymbol{\beta}_\infty,\boldsymbol{\mu}^{(n)},\mathbf{P}^{(n)})$ où l'on extrait directement l'échantillon "moyenne" $\boldsymbol{\mu}^{(n)}$ à partir de $\boldsymbol{\theta}^{(n)} = [\boldsymbol{\mu}^{(n)^T} \quad (\tilde{\boldsymbol{\sigma}}^{(n)})^T \quad (\tilde{\boldsymbol{\ell}}^{(n)})^T]^T$ tandis que $\mathbf{P}^{(n)}$ peut être déterminé en fonction de $\tilde{\boldsymbol{\sigma}}^{(n)}$ et $\tilde{\boldsymbol{\ell}}^{(n)}$ au moyen de (11). Il en résulte que $p(\Delta\boldsymbol{\beta}_\infty|\mathbf{y}_{1:M})$ peut être approché par un mélange de gaussiennes. Cependant, il faut en premier lieu générer des échantillons $\boldsymbol{\theta}^{(n)} \sim p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$, pour $n = 1\cdots N$. L'apprentissage de la loi $p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$ correspond à ce qu'on appelle la calibration. Comme on ne sait pas le faire directement, on va se tourner vers une méthode à chaîne de Markov Monte Carlo (MCMC), adaptée et efficace.

## 3.2 Expression du log-posterior et du gradient

Similairement à Heinonen (2016), on peut déduire, après développements, les expressions analytiques suivantes du log posterior $\log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$ et du gradient du log posterior $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$, en notant $\tilde{\mathbf{y}}_m \triangleq \mathbf{y}_m - (\boldsymbol{\mathcal{T}}(\mathbf{x}_m) + \boldsymbol{\mu})$, l'observation centrée, et $\tilde{\tilde{\mathbf{y}}}_m \triangleq (\mathbf{P}+\mathbf{R})^{-1}\tilde{\mathbf{y}}_m$, l'observation centrée réduite.

**Log posterior** $\log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$

$$\log p(\boldsymbol{\theta}|\mathbf{y}_{1:M}) \;=\; -\frac{1}{2}\sum_{m=1}^{M} \tilde{\mathbf{y}}_m^T(\mathbf{P}+\mathbf{R})^{-1}\tilde{\mathbf{y}}_m - \frac{1}{2}M\log\det(\mathbf{P}+\mathbf{R}) \tag{21}$$

$$+ \quad \log p(\boldsymbol{\mu}|\beta_\mu,\gamma_\mu) + \log p(\tilde{\boldsymbol{\sigma}}|\alpha_\sigma,\beta_\sigma,\gamma_\sigma) + \log p(\tilde{\boldsymbol{\ell}}|\alpha_\ell,\beta_\ell,\gamma_\ell) + C \tag{22}$$

**Gradient du log posterior** $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$

$$\frac{\partial \log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})}{\partial \boldsymbol{\mu}} \;=\; \sum_{m=1}^{M} \tilde{\tilde{\mathbf{y}}}_m - (\mathbf{P}^{SE}_{\beta_\mu,\gamma_\mu})^{-1}\boldsymbol{\mu} \tag{23}$$

$$\frac{\partial \log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})}{\partial \tilde{\boldsymbol{\sigma}}} \;=\; 2\mathrm{diag}\left(\left[\sum_{m=1}^{M} \tilde{\tilde{\mathbf{y}}}_m\tilde{\tilde{\mathbf{y}}}_m^T - M(\mathbf{P}+\mathbf{R})^{-1}\right]\mathbf{P}\right)$$
$$-(\mathbf{P}^{SE}_{\beta_\sigma,\gamma_\sigma})^{-1}(\tilde{\boldsymbol{\sigma}} - \alpha_\sigma\cdot\mathbf{1}_K) \tag{24}$$

$$\frac{\partial \log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})}{\partial \tilde{\boldsymbol{\ell}}_i} \;=\; \frac{1}{2}\mathrm{tr}\left(\left[\sum_{m=1}^{M} \tilde{\tilde{\mathbf{y}}}_m\tilde{\tilde{\mathbf{y}}}_m^T - M(\mathbf{P}+\mathbf{R})^{-1}\right]\frac{\partial\mathbf{P}}{\partial\tilde{\boldsymbol{\ell}}_i}\right)$$
$$-\left[(\mathbf{P}^{SE}_{\beta_\ell,\gamma_\ell})^{-1}(\tilde{\boldsymbol{\ell}} - \alpha_\ell\cdot\mathbf{1}_K)\right]_i \qquad \text{pour } i = 1\cdots K \tag{25}$$

### 3.3 Echantillonneur HMC dynamique

Echantillonner $p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$ est ardu en raison de la dimension $3 \cdot K$ de l'espace qui est de l'ordre de quelques centaines. Les algorithmes MCMC basiques, tels que Metropolis-Hastings, sont inadaptés, car beaucoup trop lents. Exploitant les expressions analytiques du log-posterior $\log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})]$ et de son le gradient $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$, on se tourne vers des méthodes Monte Carlo Hamiltonien (HMC), décrites par Neal (2011). Ces méthodes vont autoriser une exploration plus rapide de l'espace. Initialement connu sous le terme "Hybrid Monte Carlo" et appliqué en dynamique moléculaire, HMC est populaire dans la communauté de l'apprentissage automatique. Il est fondé sur la mécanique hamiltonienne, une branche de la mécanique classique, et la géométrie différentielle afin de simuler la dynamique d'un système physique. Cela permet à l'algorithme d'explorer plus efficacement l'espace des paramètres et de fournir dans un schéma de type Metropolis-Hastings une proposition distante et pertinente, c'est-à-dire avec un faible taux de rejet. Cela conduit à une convergence plus rapide et à un meilleur mélange de la chaîne de Markov, notamment lorsque la dimension est élevée. Toutefois, certains paramètres de contrôle de l'algorithme HMC s'avèrent délicats à régler.

Pour surmonter cette difficulté, Hoffman (2014) a proposé un échantillonneur HMC dit "dynamique", dénommé No U-Turn Sampler (NUTS). Il détermine dynamiquement les longueurs successives $L$ des pas, en utilisant une heuristique pour éviter les retours en arrière (No U-Turn), effectuant ainsi une exploration plus efficace de l'espace de recherche. Quant au pas d'intégration $\varepsilon$, il est réglé lors de la phase de "chauffe" par une méthode adaptative, dite de "dual averaging". Nous exploitons dans le papier une version originelle développée par Matthew D. Hoffman en Matlab©[1]. D'autres versions plus avancées sont au cœur de langages de programmation probabiliste, tels que *Stan*, *PyMC3* et *Turing*,

**Synthèse** En entrée, on dispose des courbes $\mathbf{y}_{1:M}$ de la figure 1, auxquelles il faut adjoindre les incertitudes de mesure associées, via la donnée de la matrice de covariance $\mathbf{R}$. Une fois spécifiée la distribution *a priori* $p(\boldsymbol{\theta})$, l'échantillonneur HMC dynamique produit en régime stationnaire des échantillons $\boldsymbol{\theta}^{(n)} \sim p(\boldsymbol{\theta}|\mathbf{y}_{1:M})$, pour $n = 1 \cdots N$. Un des intérêts de l'échantillonneur HMC est que ces échantillons sont faiblement corrélés. La distribution prédictive $p(\Delta\boldsymbol{\beta}_\infty|\mathbf{y}_{1:M})$ est alors échantillonné à partir du modèle de mélange gaussien.

## 4 Résultats

Le tableau 4 fait la synthèse des hypothèses et paramètres de fonctionnement. En cohérence avec les données de la figure 1, le modèle génératif hiérarchique produit des échantillons *a posteriori*, tels que représentés sur la figure 5. On retrouve la non-stationnarité en fonction de l'altitude, avec des fluctuations rapides au dessus de 30 km et plus lentes en dessous. Si les courbes produites s'écartent parfois des données, cela témoigne de la masse de probabilité présente dans la queue de distribution. Finalement, l'approche est apte à générer une grande variété de courbes, qui sont autant de réalisations possibles au vu des données ob-

[1] distribué par la société The MathWorks (voir le site www.mathworks.com).

servées et considérant l'*a priori* $p(\boldsymbol{\theta})$. Les échantillons générés, faiblement corrélés, peuvent être ensuite exploités pour prédire l'évolution aérodynamique et la zone d'atterrissage, via l'approximation de Monte Carlo d'intégrales multidimensionnelles en grande dimension.

| | |
|---|---|
| Données | Courbes de la figure 1, $K = 54 \quad (m = 1 \cdots 5)$<br>$\mathbf{R} = 0.3^2 \cdot \mathbf{I}$, soit des erreurs non corrélées de $\pm 1\%(3\sigma)$ |
| *A priori* | Densité $p(\boldsymbol{\mu}) : \beta_\mu = 5, \gamma_\mu = 5$<br>Densité $p(\tilde{\boldsymbol{\sigma}}) : \alpha_\sigma = 0.5, \beta_\sigma = 0.25, \gamma_\sigma = 0.5$<br>Densité $p(\tilde{\boldsymbol{\ell}}) : \alpha_\ell = 0.5, \beta_\ell = 0.5, \gamma_\ell = 0.5$ |
| NUTS | Nombre d'itérations : $N = 10000$<br>Initialisation : $\boldsymbol{\theta}_0 = [0 \cdot \mathbf{1}_K^T \quad \alpha_\sigma \cdot \mathbf{1}_K^T \quad \alpha_\ell \cdot \mathbf{1}_K^T]^T$ (*a priori* moyen)<br>Phase de chauffe : $N_0 = 1000/100$ (dual averaging)<br>Taux d'acceptation cible : $\tau = 60\%$<br>Profondeur maximale de l'arbre binaire : $N_{depth} = 10$<br>*Pas d'intégration "leapfrog" après adaptation :* $\varepsilon = 6 \cdot 10^{-4}$<br>*Nombre d'évaluations de* $\nabla \log p(\boldsymbol{\theta}\|\mathbf{y}_{1:M})$ *: 4184664*<br>*Temps de calcul mono-chaîne* $\sim 12$ *h* |

Figure 4: Données et fonctionnement de l'échantillonneur HMC

# 5 Conclusion et perspectives

En rentrée atmosphérique, la prédiction de l'évolution aérodynamique du véhicule peut être menée au moyen d'un modèle génératif hiérarchique opérant sur un faible nombre de données. La distribution prédictive peut être échantillonnée via un algorithme de Monte Carlo Hamiltonien dynamique, fondé sur la géométrie différentielle.

En perspectives, on peut évoquer l'adaptation à des données manquantes ou l'exploitation intensive de chaînes de Markov en parallèle pour accélérer le calcul. Un axe particulièrement pertinent serait d'introduire un modèle dynamique qui pourrait tenir compte de la non-stationnarité éventuelle selon le véhicule. Cela permettrait de tenir compte d'un nombre supérieur d'expériences de rentrée atmosphérique et ainsi d'améliorer encore la prédiction.

# Bibliographie


Duane, S. et al (1987), Hybrid Monte Carlo, *Physics letters B*, 195, pp. 216-222.

Gibbs, M. N. (1998), Bayesian Gaussian processes for regression and classification, Doctoral dissertation, University of Cambridge.

Heinonen M. et al (2016), Non-stationary Gaussian process regression with Hamiltonian Monte Carlo, *Artificial Intelligence and Statistics*, pp. 732-740.

Kennedy M. C. and O'Hagan A. (2001), Bayesian calibration of computer models, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 63, pp. 425-464.

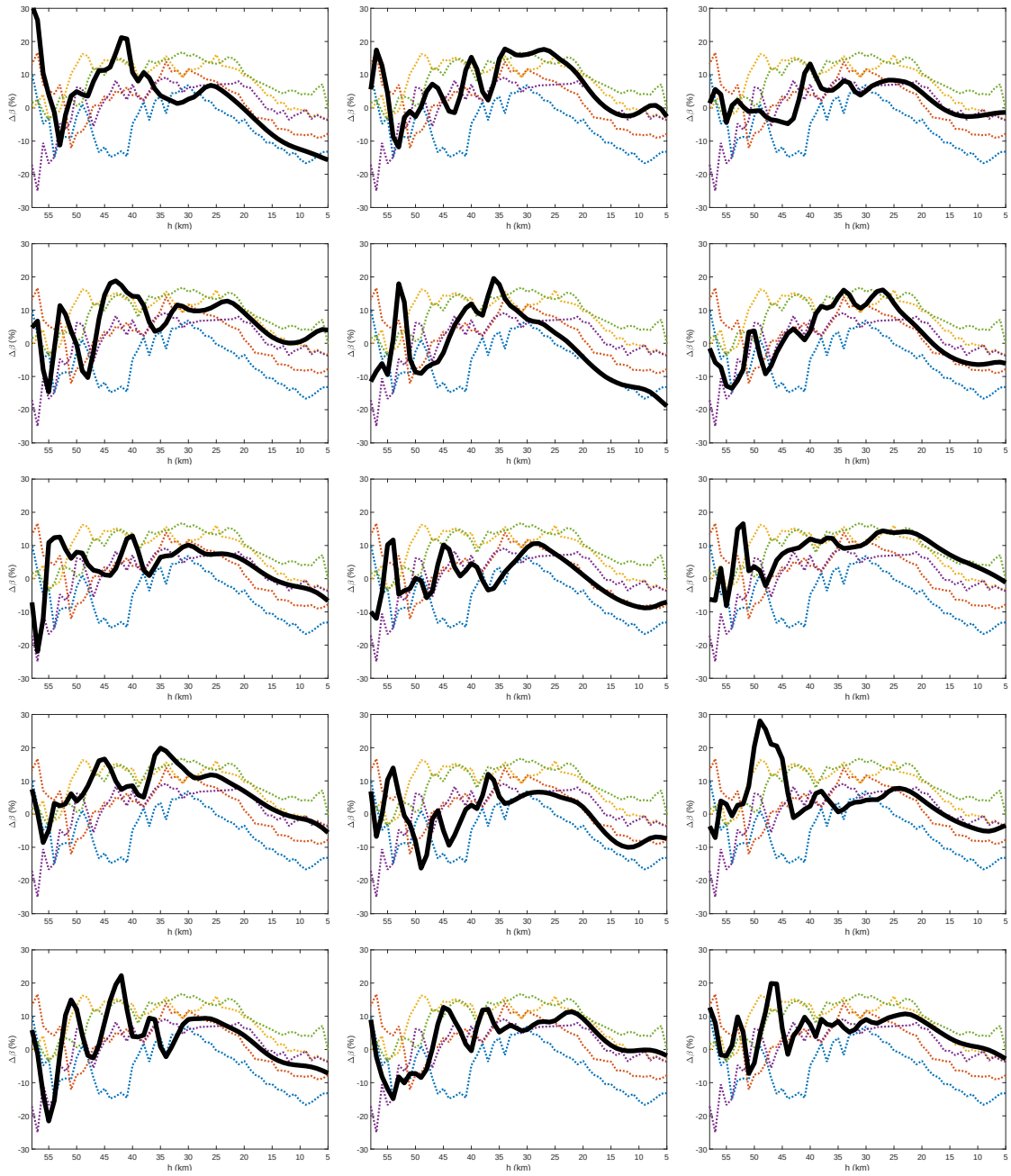



Figure 5: Echantillons *a posteriori* (trait épais noir) et originels (trait fin pointillé)

Hoffman M. D. and Gelman A. (2014), The No-U-Turn sampler : adaptively setting path lengths in Hamiltonian Monte Carlo, *J. Mach. Learn. Res.*, 15, pp. 1593-1623.

Neal R. M. (2011), MCMC using Hamiltonian dynamics, *Handbook of Markov Chain Monte Carlo*, 2., pp. 2.

Paciorek, C. et Schervish, M. (2003), Nonstationary covariance functions for Gaussian process regression, *Advances in Neural Information Processing Systems*, 16. ISO 690

# Session groupe Risques AEF

# On some depth based risk measurement for high risks

Sara Armaut*[1]

[1]Université Côte d'Azur – LJAD – France

**Résumé**

An important problem in risk theory is to understand the behavior of an expected cost associated to d $\geq$ 1 risk factors which are heterogeneous in nature. We proposed in a recent work, a depth-based Covariate-Conditional- Tail-Expectation (CCTE) in order to quantify a loss knowing that a given risk scenario occurred: considering the level sets of a depth as risk regions allows to define a direction-free CCTE. In the latter paper, we proposed an estimator of a depth-based CCTE and derived consistency results for fixed levels of risk. In a new study, we analyze the asymptotic behavior of this estimator as the risk level decreases, meaning that we study consistency of this risk measure for high risks.

**Mots-Clés:** depth, high risks, CCTE

*Intervenant

# Robust estimation of discrete distributions under local differential privacy

Flore Sentenac*[1] and Julien Chhor*†

[1]HEC – Centre de Recherche en Économie et STatistique (CREST) – France

**Résumé**

In the realms of insurance and finance, minimizing the risk of model failure is of paramount importance. Robust learning methodologies have been extensively explored in these domains. More recently, spurred by regulatory requirements and societal considerations, significant emphasis has been placed on ensuring that algorithms prioritize user privacy.
Although robust learning and local differential privacy are both widely studied fields of research, combining the two settings is just starting to be explored. We consider the problem of estimating a discrete distribution in total variation from $n$ contaminated data batches under a local differential privacy constraint.

A fraction $1-\alpha$ of the batches contain $k$ i.i.d. samples drawn from a discrete distribution $p$ over $d$ elements. To protect the users' privacy, each of the samples is privatized using an $\epsilon$-locally differentially private mechanism. The remaining $\alpha n$ batches are an adversarial contamination. The minimax rate of estimation under contamination alone, with no privacy, is known to be $\alpha/\sqrt{k} + \sqrt{d/kn}$. Under the privacy constraint alone, the minimax rate of estimation is $\sqrt{d^2/\epsilon^2 kn}$. We show, up to a $\sqrt{\log(1/\alpha)}$ factor, that combining the two constraints leads to a minimax estimation rate of $\alpha\sqrt{d/\epsilon^2 k} + \sqrt{d^2/\epsilon^2 kn}$, larger than the sum of the two separate rates.
We provide a polynomial-time algorithm achieving this bound, as well as a matching information theoretic lower bound.

**Mots-Clés:** Privacy, Robust

*Intervenant
†Auteur correspondant: julien.chhor@ensae.fr

# OT and EOT QQ-plots. Application in Risk Analysis and Management

Marie Kratz*[1]

[1]ESSEC Business School – CREAR – France

**Résumé**

Univariate Q-Q plot is a very powerful visualisation tool, used to compare two distributions. Turning to multivariate quantiles, various approaches are possible. We refer the reader to Serfling (2002), Chaudhury (1996), Singha et al. (2023-24), Singha (2024) and references therein for an extensive survey on multivariate quantiles. Due to the absence of natural ordering, there is no straightforward extension of QQ plots for multivariate samples, as there is no natural extension of quantile function for multivariate distributions.
Using the geometric multivariate quantiles developed by Chaudhury (1996) and their property of unique characterization of the underlying distribution, Dhar et al. (2014) constructed component wise QQ plots for comparing multivariate distributions. Considering a similar approach as Easton and McCulloch (1990) and Dhar et al. (2014), Singha et al. extended this graphical tool when using optimal transport (OT) map and optimal potential (OP) function, referred as OT QQ plot and OP QQ plot, respectively. It was also shown that, as the size of the samples increases, the Q-Q plots become arbitrarily close to the straight line passing through the origin and with slope 1 if and only if the samples are drawn from the same distribution.

In order to generate an OT Q-Q plot, one must first calculate empirical OT maps. However, computing empirical OT maps can be costly, especially when size of the sample is large. Practical solutions have been proposed in the literature, one of the most popular approaches being entropy regularization (Cuturi, 2013). By selecting the regularization parameter to be sufficiently small, the entropy regularized map (EOT) can closely approximate the OT map. Moreover, the EOT also characterizes a distribution uniquely (see Singha, 2024), justifying the construction of OT and EOT QQ-plots. As for geometric quantiles (see Dhar et al., 2014), test statistics for comparing two distributions based on the proposed Q-Q plots, can also be developed to assess their relevance.

Turning to applications, we show the attractiveness of (E)OT QQ plots to develop stress scenarios for risk management purpose. This approach should provide regulators and risk managers with a robust tool to identify the most suitable extreme scenarios for targeted stress testing at any specified probability level.

This presentation is based on joint works with S. Singha (TIFR-CAM), S. Vadlamani (TIFR-CAM) and M. Dacorogna (PRS).

References:

*Intervenant

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems* 26, 2292–2300

Dhar, S., Chakraborty, B. and Chaudhuri, P. (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests. *Bernoulli* 20(3), 1484–1506

Easton, G.S. and McCulloch, R.E. (1990). A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association* 85, 376–386

Serfling, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. Statistica Neerlandica 56, 214–232

Singha, S. (2024). Characterising distributions and their tails using multivariate quantiles and depths. *Doctoral Thesis in Mathematics*, TIFR
Singha, S., Kratz, M., Vadlamani, S. (2023-24). Extremal behaviour and convergence rates for sample–based geometric quantiles and half space depths. *ArXiv:2023-0610789v2*

**Mots-Clés:** Multivariate quantiles, Optimal Transport (OT) Map, Entropy OT (EOT), QQ plots, Quantitative risk analysis, Stress scenarios, Test statistics

# Apprentissage sur données déséquilibrées

# Étude théorique et expérimentale de SMOTE : limites et comparaisons des stratégies de rééquilibrage

Abdoulaye SAKHO [1,2], Erwan SCORNET [2] & Emmanuel MALHERBE [3]

[1] *Artefact Research Center, Paris, France. abdoulaye.sakho@artefact.com*
[2] *Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France. erwan.scornet@sorbonne-universite.fr*
[3] *Artefact Research Center, Paris, France. emmanuel.malherbe@artefact.com*

**Résumé.** *Synthetic Minority Oversampling Technique* (SMOTE) est une stratégie de rééquilibrage courante pour traiter les ensembles de données déséquilibrés. Asymptotiquement, nous prouvons que SMOTE (avec la valeur par défaut de son hyperparamètre) régénère la distribution originale en copiant simplement les échantillons minoritaires initialement présents. Nous introduisons ensuite deux nouvelles stratégies liées à SMOTE et les comparons aux procédures de rééquilibrage les plus récentes. Nous montrons que les stratégies de rééquilibrage ne sont nécessaires que lorsque l'ensemble de données est fortement déséquilibré. Pour de tels ensembles de données, SMOTE, nos propositions ou les procédures de sous-échantillonnage sont les meilleures stratégies.

**Mots-clés.** Classification, Données déséquilibrées, SMOTE

**Abstract.** *Synthetic Minority Oversampling Technique* (SMOTE) is a common rebalancing strategy for handling imbalanced data sets. Asymptotically, we prove that SMOTE (with default parameter) regenerates the original distribution by simply copying the original minority samples. Then we introduce two new SMOTE-related strategies, and compare them with state-of-the-art rebalancing procedures. We show that rebalancing strategies are only required when the data set is highly imbalanced. For such data sets, SMOTE, our proposals, or undersampling procedures are the best strategies.

**Keywords.** Classification, Imbalanced data sets, SMOTE

## 1 Contexte

Les ensembles de données déséquilibrés sont un problème typique rencontré dans plusieurs applications (He and Garcia, 2009), telles que la détection de fraude (Hassan and Abraham, 2016), la prédiction de diagnostics médicaux (Khalilia et al., 2011) et même la prédiction de l'appétence au désabonnement (Nguyen and Duong, 2021). En présence d'ensembles de données déséquilibrés, la plupart des algorithmes d'apprentissage statistiques ont tendance à prédire la classe majoritaire. Plusieurs stratégies ont été développées pour traiter ce problème, comme expliqué par Krawczyk (2016) et Ramyachitra and Manikandan (2014).

SMOTE est l'algorithme central dont dérivent la plupart des algorithmes qui génèrent de nouvelles données synthétiques au sein de la classe minoritaire. En effet, à l'exception des algorithmes basés sur les réseaux de neuronnes, la plupart des autres stratégies utilisent toujours une variante de l'interpolation linéaire incluse dans la procédure SMOTE. Plusieurs variantes tentent de se concentrer sur la génération d'échantillons synthétiques proches de la limite du support de la classe minoritaire. La plus courante est *ADASYN* (He et al., 2008) dont l'idée principale est de produire davantage d'échantillons synthétiques par interpolation linéaire entre les échantillons de la classe minoritaire qui sont principalement entourés d'échantillons de la classe majoritaire. BorderLine SMOTE (Han et al., 2005) cherche à générer de nouveaux échantillons synthétiques à la frontière des deux classes. Une autre variante de SMOTE axée sur les frontières est SVM-SMOTE (Nguyen et al., 2011), dont l'idée est de commencer par appliquer un classificateur de machine à vecteur de support aux données déséquilibrées. L'interpolation linéaire est ensuite effectuée sur le vecteur de soutien de la classe minoritaire.

Il existe plusieurs travaux théoriques concernant les stratégies de rééquilibrage. La méthode de pondération par classe est étudiée théoriquement par King and Zeng (2001). King and Zeng (2001) étudient l'effet de la stratégie Random Under Sampling sur un classificateur de régression logistique. À notre connaissance, il n'existe que peu de travaux théoriques disséquant la machinerie intrinsèque de l'algorithme SMOTE. Par exemple, Elreedy and Atiya (2019) calculent l'espérance et la matrice de covariance des points générées par SMOTE. De plus, Elreedy et al. (2023) établissent une expression de la densité des échantillons générés par SMOTE à partir de la densité des échantillons originaux de la classe minoritaire.

**Notations** On note par $\mathcal{U}([a,b])$ la distribution uniforme sur le segment $[a,b]$. La distribution gaussienne multivariée centré en $\mu$ et de matrice de covariance $\Sigma$ est notée $\mathcal{N}(\mu,\Sigma)$. Pour tout ensemble $A$, on note par $Vol(A)$, la mesure de Lebesgue de $A$. Pour tout $z \in \mathbb{R}^d$ et$r > 0$, $B(z,r)$ est la boule centrée en $z$ et de rayon $r$. On note par $c_d = Vol(B(0,1))$ le volume de la boule unitaire dans $\mathbb{R}^d$. Pour tout $p,q \in \mathbb{N}$, et tout $z \in [0,1]$, on note $\mathcal{B}(p,q;z) = \int_{t=0}^{z} t^{p-1}(1-t)^{q-1}\mathrm{d}t$ la fonction beta incomplète.

# 2 Étude de SMOTE

Dans cette section, nous étudions l'algorithme SMOTE, qui génère des données synthétiques par interpolation linéaire entre deux instances originales de la classe minoritaire. L'algorithme SMOTE possède un seul hyperparamètre, $K$, qui représente le nombre de plus proches voisins pris en compte lors de l'interpolation. Une seule itération SMOTE est détaillée dans Algorithm 1. Dans un pipeline classique d'apprentissage statistique, une itération de SMOTE est répétée autant de fois que nécessaire afin d'obtenir un ratio prédéfini entre les deux classes avant d'entraîner un classifieur.

Il a été démontré que SMOTE présentait de bonnes performances lorsqu'il était associé à des algorithmes de classification (voir par exemple Mohammed et al., 2020). Nous supposons

que $X_1, \ldots, X_n$ sont des échantillons indépendants et identiquement distribuées de la classe minoritaire (c'est-à-dire $Y_i = 1$ pour tous $i \in [n]$), avec une densité commune $f_X$ à support borné, dénotée par $\mathcal{X}$. Enfin, on note par $f_{Z_{K,n}}$, la densité de SMOTE associé aux paramètres $K, n$ appliquée sur les échantillons de $X_1, \ldots, X_n$ (pour tout $K, n$).

**Algorithm 1** Une itération de SMOTE.

**Input:** Échantillons de la classe minoritaire $X_1, \ldots, X_n$, nombre $K$ de plus proches voisins.
Choisir uniformément $X_c$ (appellé **point central**) parmi $\{X_1, \ldots, X_n\}$.
On note $I = X_{(1)}(X_c, ) \ldots, X_{(K)}(X_c)$, les $K$ plus proches voisins de $X_c$ (norme $L_2$ ).
Sélectionner $X_k \in I$ uniformément.
$w \leftarrow \mathcal{U}([0, 1])$
$Z \leftarrow X_c + w(X_k - X_c)$
**Return** $Z$

Nous nous plaçons dans le cadre de variables d'entrée continues, puisque les procédures synthétiques telles que SMOTE sont conçues à l'origine pour traiter de telles variables.

**Théorème 2.1.** *Supposons qu'il esixte $R > 0$, tel que $\mathcal{X} \subset B(0, R)$. De plus, supposons qu'il existe $C_2$ tel que, pour tout $x \in \mathbb{R}^d$, $f_X(x) \leq C_2 \mathbb{1}_{x \in \mathcal{X}}$. Alors, pour tout $n \geq K \geq 1$, pour tout $x_c \in \mathcal{X}$ et pour tout $\alpha > 0$, on a*

$$\mathbb{P}(|Z_{K,n} - X_c| \geq \alpha | X_c = x_c) \leq \varepsilon(n, \alpha, K, x_c), \tag{1}$$

*où*

$$\varepsilon(n, K, x_c, \alpha) = c_d R^d \eta(\alpha, R) \exp\left[ n \left( 3 \sqrt{\frac{e}{1 - \beta_{x_c,\alpha}}} \sqrt{\frac{K}{n}} + \ln\left(1 - \beta_{x_c,\alpha}\right) \right) \right], \tag{2}$$

*avec $\beta_{x_c,\alpha} = \mu_X\left(B\left(x_c, \alpha\right)\right) > 0$ et*

$$\eta(\alpha, R) = \begin{cases} C_2 \ln\left(\frac{2R}{\alpha}\right) & \text{if } d = 1, \\ \frac{C_2}{d-1}\left(\left(\frac{2R}{\alpha}\right)^{d-1} - 1\right) & \text{if } d > 1, \\ 0 & \text{if } \alpha > 2R. \end{cases}$$

*Par conséquent, si $\lim_{n\to\infty} K/n = 0$, on a alors, pour tout $x_c \in \mathcal{X}$, $Z_{K,n}|X_c = x_c \to x_c$ en probabilité.*

Nous montrons tout d'abord que la densité des points générés par la procédure SMOTE, avec la valeur par défaut pour $K$, converge en probalité vers la densité de la classe minoritaire, lorsque le nombre d'échantillons minoritaires augmente. Nous prouvons (Theorem 2.1) également que, sans réglage de l'hyperparamètre $K$ (habituellement fixé à 5), SMOTE copie asymptotiquement les échantillons minoritaires originaux, manquant ainsi de la variabilité intrinsèque désirée dans toute procédure générative synthétique. Cela souligne l'importance du réglage de l'hyperparamètre dans SMOTE, lorsque le nombre d'échantillons de la classe minoritaire est suffisamment grand.

# 3 Nouvelles stratégies

Les limites de SMOTE mises en évidence dans la section 2 nous conduisent à deux nouvelles stratégies de rééquilibrage.

**CV SMOTE** Nous introduisons un nouvel algorithme, appelé CV SMOTE, qui trouve le meilleur hyperparamètre $K$ parmi une grille prédéfinie via une procédure de validation croisée 5-fold. La grille est composée de l'ensemble $\{1, 2, \ldots, 15\}$ étendu avec les valeurs $\lfloor 0.01 n_{train} \rfloor$, $\lfloor 0.1 n_{train} \rfloor$ et $\lfloor \sqrt{n_{train}} \rfloor$, où $n_{train}$ est le nombre d'échantillons minoritaires dans l'ensemble d'apprentissage.

Rappelons qu'à travers Theorem 2.1, nous montrons que la procédure SMOTE avec la valeur par défaut de l'hyperparamètre $K = 5$ copie asymptotiquement les échantillons originaux. L'idée de CV SMOTE est donc d'essayer plusieurs valeurs de $K$ afin d'éviter de copier les échantillons et d'obtenir probablement une amélioration de la performance prédictive du classificateur utilisé par la suite.

**Multivariate Gaussian SMOTE($K$)** Nous introduisons maintenant une nouvelle stratégie de suréchantillonnage que nous nommons Multivariate Gaussian SMOTE (MGS). Dans cette procédure, nous générons de nouveaux échantillons à partir de la distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$, où la moyenne empirique et la matrice de covariance ($\hat{\mu}$ et $\hat{\Sigma}$ respectivement) sont estimées à l'aide des $K$ plus proches voisins et du point central. Nous détaillons une itération MGS dans l'algorithme 2. L'idée sous-jacente de MGS est d'exploiter au maximum le voisinage du point central. L'utilisation d'une distribution gaussienne multivariée, dont le support n'est pas borné, réduit le risque de copier simplement les échantillons originaux lorsque $K/n \to 0$.

**Algorithm 2** Une itération de Multivariate Gaussian SMOTE.

**Input:** Échantillons de la classe minoritaire $X_1, \ldots, X_n$, nombre de plus proches voisins $K$.
Choisir uniformément $X_c$ (appellé **point central**) parmi $\{X_1, \ldots, X_n\}$.
On note $I = X_{(1)}(X_c, ) \ldots, X_{(K)}(X_c)$, les $K$ plus proches voisins de $X_c$ (norme $L_2$ ).
$\hat{\mu}(X_c) \leftarrow \frac{1}{K+1} \sum_{X_k \in I \cup \{X_c\}} X_k$
$\hat{\Sigma}(X_c) \leftarrow \frac{1}{K+1} \sum_{X_k \in I \cup \{X_c\}} (X_k - \hat{\mu})^T (X_k - \hat{\mu})$
Sample $Z \sim \mathcal{N}\left(\hat{\mu}(x_c), \hat{\Sigma}(x_c)\right)$
**Return** $Z$

# 4 Résultats

**Protocole** Nous comparons les différentes stratégies de rééquilibrage sur 11 ensembles de données réelles décrits dans la Table 1. Nous utilisons un stratified-split de 80%/20% (formation/test) des données, et appliquons chaque stratégie de rééquilibrage sur l'ensemble

Table 1: Description des ensembles de données, où $n$ est le nombre d'échantillons et $d$ le nombre de variables explicatives.

| | Total $N$ | minority samples $n/N$ | $d$ |
|---|---|---|---|
| GA4 | 319 066 | 0.7% | 7 |
| Credit | 284 315 | 0.2% | 29 |
| Abalone | 4 177 | 1% | 8 |
| Phoneme | 5 404 | 29% | 5 |
| Yeast | 1 462 | 11% | 8 |
| Pima | 768 | 35% | 8 |
| Wine | 4 974 | 4% | 11 |
| Vehicule | 846 | 23% | 18 |
| Ionosphere | 351 | 36% | 32 |
| Haberman | 306 | 26% | 3 |
| Breast cancer | 630 | 36% | 9 |

d'entrainement, afin d'obtenir un ensemble de données équilibré. Une procédure d'apprentissage (régression logistique ou Random Forest) avec des hyperparamètres par défaut est entraînée sur l'ensemble d'entraînement rééquilibré. La performance est évaluée sur l'ensemble de test via le ROC AUC. Cette procédure est répétée 100 fois et la moyenne des résultats est calculée. Nous utilisons l'implémentation de *imb-learn* (Lemaître et al., 2017) pour les stratégies état de l'art.

**Méthodes de rééquilibrage** Notons que tous les ensembles de données présentés dans Table 2 sont fortement déséquilibrés, avec un ratio inférieur à 1% (10% et 10% pour les ensembles de données Pima et Haberman(10%) respectivement). Alors que dans la grande majorité des scénarios, None fait partie des meilleures approches pour traiter les données déséquilibrées, elle semble être surpassée par les stratégies de rééquilibrage dédiées aux ensembles de données fortement déséquilibrés, présentées dans la Table 2. Par conséquent, en ne considérant que les variables d'entrée continue et en mesurant la performance prédictive avec l'AUC ROC, nous observons qu'appliquer une stratégie de rééquilibrage n'est nécessaire que dans un cas précis : la classe minoritaire est fortement sous-représentée. En outre, nous constatons que la stratégie RUS présente une meilleure amélioration des performances du classifier pour les très grands ensembles de données, qui devraient être moins sensibles à la perte d'informations due au sous-échantillonnage. Plusieurs articles précurseurs avaient déjà remarqué que la stratégie None était compétitive en termes de performances prédictives. He et al. (2008) comparent la stratégie None, ADASYN et SMOTE, avant l'entraînement d'un d'arbre de décision sur des ensembles de données de 5 (comprenant Vehicle, Pima, Ionosphere et Abalone). En termes de précision et de score F1, la stratégie None est à égalité avec les deux autres méthodes de rééquilibrage. Han et al. (2005) étudient l'impact de Borderline SMOTE et d'autres variantes de SMOTE sur 4 ensemble de données (y compris Pima et Haberman). La stratégie None est compétitive (en termes de score F1) sur deux de ces ensembles de données.

Table 2: ROC AUC pour Random Forest pour différentes stratégies de rééquilibrage et différents ensembles de données. Seuls les ensembles de données pour lesquels la stratégie None ne figure pas parmi les meilleures (en gras) sont affichés. Les ensembles de données artificiellement sous-échantillonnés au niveau de la classe minoritaire sont en italique.

| Resampling Strategy | None | Class weight | RUS | ROS | Near Miss1 | BS1 | Smote | CV Smote | MGS |
|---|---|---|---|---|---|---|---|---|---|
| GA4 (1%) | 0.660 | 0.472 | **0.866** | 0.500 | 0.848 | 0.652 | 0.506 | 0.720 | 0.650 |
| Credit (0.2%) | 0.939 | 0.938 | **0.975** | 0.941 | 0.906 | 0.945 | 0.954 | 0.954 | 0.950 |
| Abalone (1%) | 0.697 | 0.702 | 0.719 | 0.712 | 0.570 | 0.712 | 0.756 | 0.750 | **0.799** |
| *Phoneme* (1%) | 0.819 | 0.821 | 0.851 | 0.814 | 0.575 | 0.847 | 0.876 | 0.877 | **0.899** |
| *Yeast* (1%) | 0.906 | 0.928 | 0.931 | 0.929 | 0.806 | 0.946 | 0.967 | **0.968** | 0.944 |
| Wine (4%) | 0.819 | 0.815 | **0.846** | 0.810 | 0.748 | 0.827 | 0.828 | 0.822 | 0.822 |
| *Pima* (10%) | 0.797 | 0.804 | 0.802 | 0.800 | 0.680 | 0.812 | 0.807 | 0.806 | **0.821** |
| *Haberman* (10%) | 0.580 | 0.580 | 0.599 | 0.582 | **0.634** | 0.609 | 0.617 | 0.598 | 0.619 |

**SMOTE** Nous remarquons que les performances de CV SMOTE sont comparables à celles de SMOTE avec l'hyperparamètre par défaut ($K = 5$). Cela peut s'expliquer par le choix de notre grille (qui pourrait être étendue) ou par les caractéristiques de l'ensemble des données. En effet, le seul jeu de données pour lequel nous constatons que CV SMOTE est notablement meilleur que SMOTE est GA4, qui contient le plus grand nombre d'échantillons minoritaires. Cela correspond à notre analyse théorique (Theorem 2.1) qui souligne que SMOTE, par défaut, tend à copier les échantillons minoritaires originaux, lorsque le nombre d'échantillons minoritaires est suffisamment important. Il conviendrait donc d'effectuer d'autres analyses pour étudier l'efficacité potentielle de CV SMOTE lorsque le nombre d'échantillons minoritaires est suffisamment élevé.

**MGS** Cette nouvelle stratégie présente de bonnes améliorations des performances prédictives. En effet, comme le montre la Table 2, MGS présente la meilleure amélioration sur 3 ensembles de données. Cela pourrait s'expliquer par l'échantillonnage gaussien des observations synthétiques qui permet aux points de données générés de se situer en dehors de l'enveloppe convexe de la classe minoritaire. MGS est potentiellement une nouvelle stratégie prometteuse, qui sera disponible sous la forme d'un logiciel libre.

# 5 Conclusion

Notre travail dans cet article est à la fois théorique et expérimental. Nous avons d'abord prouvé que SMOTE (avec le paramètre par défaut) régénère la distribution originale en copiant simplement les échantillons minoritaires originaux. Ces résultats théoriques nous ont permis d'introduire deux nouvelles stratégies permettant de générer des instances synthétiques au sein de la classe minoritaire.

D'autres expériences devraient être menées pour comprendre les performances surprenantes de RUS, qui surpasse systématiquement ROS, alors que les deux méthodes très similaires, car elles reposent toutes deux sur le rééchantillonnage. Enfin, afin d'analyser MGS(K) plus en détail, nous aimerions étudier l'impact d'un facteur de renormalisation $\lambda$ dans l'estimation de la matrice de covariance, de sorte que la dernière étape de Algorithm 2 deviendrait $Z \sim \mathcal{N}(\hat{\mu}, \lambda\hat{\Sigma})$.

# References


Elreedy, D. and A. F. Atiya (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences 505*, 32–64.

Elreedy, D., A. F. Atiya, and F. Kamalov (2023, January). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*.

Han, H., W.-Y. Wang, and B.-H. Mao (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer.

Hassan, A. K. I. and A. Abraham (2016). Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*, pp. 117–127. Springer.

He, H., Y. Bai, E. A. Garcia, and S. Li (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328. Ieee.

He, H. and E. A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering 21*(9), 1263–1284.

Khalilia, M., S. Chakraborty, and M. Popescu (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making 11*, 1–13.

King, G. and L. Zeng (2001). Logistic regression in rare events data. *Political Analysis 9*(2), 137–163.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence 5*(4), 221–232.

Lemaître, G., F. Nogueira, and C. K. Aridas (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research 18*(17), 1–5.

Mohammed, A. J., M. M. Hassan, and D. H. Kadir (2020). Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering 9*(3), 3161–3172.

Nguyen, H. M., E. W. Cooper, and K. Kamei (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms 3*(1), 4–21.

Nguyen, N. N. and A. T. Duong (2021). Comparison of two main approaches for handling imbalanced data in churn prediction problem. *Journal of advances in information technology 12*(1).

Ramyachitra, D. and P. Manikandan (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR) 5*(4), 1–29.

# Semantic segmentation of forest point clouds using neural network

Yuchen Bai [1] Jean-Baptiste Durand [2] Grégoire Vincent [3] Florence Forbes [4]

[1] *UGA, France, yuchen.bai@inria.fr*
[2] *CIRAD, France, Jean-Baptiste.Durand@cirad.fr*
[3] *IRD, France, gregoire.vincent@ird.fr*
[4] *INRIA, France, florence.forbes@inria.fr*

**Résumé.** Depuis des années, la technologie LiDAR (Light Detection And Ranging) s'est imposée comme un outil indispensable pour acquérir des données 3D approfondies dans le domaine de la gestion forestière. Aussi connue sous le nom de scanner laser, cette technologie permet l'acquisition de données sous forme de nuages de points, offrant des détails minutieux sur la structure de la canopée. En particulier, le LiDAR offre la possibilité d'établir des modèles foliaires de forêts avec une précision sans précédent, avec pour motivation le rôle crucial joué par la surface foliaire dans les échanges gazeux entre la végétation et l'atmosphère. La surveillance de la surface foliaire en forêt contribue de manière significative à la compréhension des flux saisonniers dans les forêts tropicales, améliorant ainsi la précision des modèles climatiques pour prédire les impacts du réchauffement.

Divers véhicules sont employés pour l'acquisition de données, avec la numérisation laser terrestre (TLS), qui fournit des données 3D détaillées mais nécessite d'intensives interventions sur le site. La numérisation laser aéroportée (ALS) couvre des zones plus vastes mais avec une densité de points plus faible, source de défis pour l'observation de la végétation de sous-étage, en raison des occlusions causées par la canopée. Les avantages relatifs de la numérisation laser par drone (ULS) se manifestent dans ce contexte, en offrant la possibilité d'une collecte de données dense mais sans nécessité d'intervention sur site. L'enjeu clé de cette étude est l'obtention d'une segmentation sémantique précise pour distinguer les feuilles du bois, exigée en monitoring des forêts pour surveiller les variations de densité foliaire. Ses applications comprennent la prédiction de la séquestration du carbone, la surveillance des maladies et la planification de la récolte. Les méthodes existantes pour les données TLS rencontrent des difficultés lorsqu'elles sont appliquées à l'ULS en raison d'un déséquilibre entre les deux classes et de la dépendance à des informations peu fiables. Afin de résoudre ce problème, nous proposons une nouvelle approche nommée SOUL (Semantic segmentation On ULs) [1], utilisant uniquement les coordonnées des points en entrée du réseau neuronal pour garantir son adaptabilité à différentes localisations et avec divers capteurs.

L'apport de ce travail est triple. Tout d'abord, SOUL est la première approche conçue pour la segmentation sémantique sur des nuages de points ULS dans les forêts tropicales, démontrant une meilleure classification des points de bois. Ensuite, la méthode de prétraitement des données GVD (Geodesic Voxelization Decomposition) relève le défi de l'entraînement des réseaux neuronaux à partir de nuages de points épars dans des environnements forestiers tropicaux. Enfin, nous proposons une fonction de perte ré-équilibrée comme solution polyvalente pour résoudre le problème de déséquilibre des classes dans diverses architectures

d'apprentissage profond. Notre étude offre des perspectives prometteuses pour l'application de drones en monitoring des forêts, en repoussant les limitations des méthodologies existantes et en contribuant à une amélioration de la segmentation sémantique dans des environnements forestiers tropicaux.

**Mots-clés.** Apprentissage profond, segmentation sémantique, déséquilibre de classe, LiDAR, nuage de points

**Abstract.** In recent decades, LiDAR technology has become an indispensable tool for collecting extensive 3D data in the field of forest inventory. Often known as laser scanning, this technology facilitates the acquisition of point cloud data, providing detailed insights into canopy structure. In particular, LiDAR provides the opportunity to map forest leaf area with unprecedented accuracy, while leaf area has remained an important source of uncertainty affecting models of gas exchanges between the vegetation and the atmosphere. The vigilant monitoring of leaf area contributes significantly to comprehending the seasonal fluxes in tropical forests, thereby refining the precision of climate models in predicting the repercussions of global warming.

Various vehicles are used for data collection, with terrestrial laser scanning (TLS) providing detailed but labor-intensive 3D data. Airborne laser scanning (ALS) covers larger areas but with lower point density, posing challenges for observing understory vegetation due to canopy occlusions. The rise of UAV laser scanning (ULS) addresses these challenges, offering dense data collection without on-site intervention. In fact, forest monitoring requires accurate semantic segmentation to distinguish leaves from wood, which is crucial for monitoring foliage density variations with applications in carbon sequestration, disease monitoring, and harvest planning. Existing methods for TLS data face challenges when applied to ULS due to class imbalance and reliance on unreliable intensity information. To address this, we propose an end-to-end approach named SOUL (Semantic segmentation On ULs) [1] utilizing only point coordinates as input of the neural network for versatility across locations and sensors.

The contributions of this work are three-fold. First, SOUL is the first approach designed for semantic segmentation on ULS point clouds in tropical forests, showcasing superior wood point classification. Second, the GVD (Geodesic Voxelization Decomposition) preprocessing method addresses the challenge of training neural networks from sparse point clouds in tropical forest environments. Third, the proposed rebalanced loss function provides a versatile solution for addressing class imbalance in various deep learning architectures. Our research offers promising insights into the application of ULS for forest monitoring, bridging gaps in existing methodologies and laying the foundation for improved semantic segmentation in challenging tropical forest environments.

**Keywords.** Deep Learning, Semantic Segmentation, Class Imbalance, LiDAR, Point Cloud

# 1 Introduction

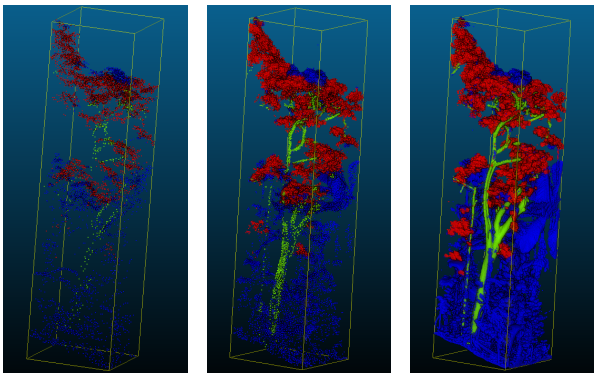
In the past decades, LiDAR technology has been frequently used to acquire massive 3D data in the field of forest inventory (Vincent et al. [2]; Ullrich & Pfennigbauer [3]). The acquisition of point cloud data by employing LiDAR technology is referred to as laser scanning. The collected point cloud data provides rich details on canopy structure, allowing us to calculate a key variable, leaf area, which controls water efflux and carbon influx. Monitoring leaf area should help in better understanding processes underlying flux seasonality in tropical forests, and is expected to enhance the precision of climate models for predicting the effects of global warming. There are various types of vehicles for data collection, with ground-based equipment and aircraft being the most commonly employed. The former operates a bottom-up scanning called terrestrial laser scanning (TLS), providing highly detailed and accurate 3D data. Scans are often acquired in a grid pattern every 10 m and co-registered into a single point cloud. However, TLS requires human intervention within the forest, which is laborious and limits its extensive implementation. Conversely, airborne laser scanning (ALS) is much faster and can cover much larger areas. Nonetheless, the achieved point density is typically two orders of magnitude smaller due to the combined effect of high flight altitude and fast movement of the sensor. Additionally, occlusions caused by the upper tree canopy make it more difficult to observe the understory vegetation.

In recent years, the development of drone technology and the decreasing cost have led to UAV laser scanning (ULS) becoming one favored option (Brede et al. [4]). It does not require in-situ intervention and each flight can be programmed to cover a few hectares. The acquired data is much denser than ALS (see Figure 1(a) and Figure 1(b)), which provides us with more comprehensive spatial information. Increasing the flight line overlap results in multiple angular sampling, higher point density and mitigates occlusions. Although the data density is still relatively low, compared with TLS, ULS can provide previously unseen overstory details due to the top-down view and overlap flight line. Furthermore, ULS is considered to be more suitable for conducting long-term monitoring of forests than TLS, as it allows predefined flight plans with minimal operator involvement.

Consequently, leaf-wood semantic segmentation in ULS data is required to accurately monitor foliage density variation over space and time. Changes in forest foliage density are indicative of forest functioning and their tracking may have multiple applications for carbon sequestration prediction, forest disease monitoring and harvest planning. Fulfilling these requirements necessitates the development of a robust algorithm that is capable to classify leaf and wood in forest environments. While numerous methods have demonstrated effective results on TLS data, these methods cannot be applied directly to ULS, due in particular to the class imbalance issue: leaf points account for about 95% of the data. Another problem is that many methods rely on the extra information provided by LiDAR devices, such as intensity. In the context of forest monitoring, intensity is not reliable due to frequent pulse fragmentation and variability in natural surface reflectivity (see Vincent et al. [5]). Furthermore, the reflectivity of the vegetation is itself affected by diurnal or seasonal changes in physical conditions, such as water content or leaf orientation (Brede et al. [4]). Therefore, methods relying on intensity information (Wu et al. [6]) may exhibit substantial variations in performance across different locations and even within the same location for different

acquisition batches. To address this issue, certain methods (LeWos proposed by Wang et al. [7]; Morel et al. [8]) have good results while exclusively utilizing the spatial coordinates of LiDAR data.

Inspired by the existing methods, we propose a novel end-to-end approach **SOUL** (Semantic segmentation On ULs) based on PointNet++ proposed by Qi et al. [9] to perform semantic segmentation on ULS data. SOUL uses only point coordinates as input, aiming to be applicable to point clouds collected in forests from various locations worldwide and with sensors operating at different wavelengths. The foremost concern to be tackled is the acquisition of labeled ULS data. Since no such data set existed up to now, we gathered a ULS data set comprising 282 trees labeled. This was achieved through semi-automatic segmentation of a coincident TLS point cloud and wood/leaf label transfer to ULS point cloud. Secondly, the complex nature of tropical forests necessitates the adoption of a data pre-partitioning scheme. While certain methods (Krisanski et al. [10]; Wu et al. [11]) employ coarse voxels with overlap, such an approach leads to a fragmented representation and incomplete preservation of the underlying geometric information. The heterogeneous distribution of points within each voxel, including points from different trees and clusters at voxel boundaries, poses difficulties for data standardization. We introduce a novel data preprocessing methodology named geodesic voxelization decomposition (GVD), which leverages geodesic distance as a metric for partitioning the forest data into components and uses the topological features, like intrinsic-extrinsic ratio (IER) (He et al. [12]; Liu et al. [13]), to preserve the underlying geometric features at component level (see Section **??**). The last issue concerns the class imbalance problem during the training stage. To address this issue, we developed a novel loss function named the rebalanced loss, which yielded improved performance compared with the focal loss (Lin et al. [14]) for our specific task. This enhancement resulted in a 23% increase in the ability to recognize wood points, see Table 3.



(a) ALS (b) ULS (c) TLS

Figure 1: Point clouds produced by three scanning modes on the same area (20m × 20m × 42m), illustrate how much the visibility of the understory differs. The colors in the figure correspond to different labels assigned to the points, where red and green indicate leaves and wood, respectively. Blue points are unprocessed, so labeled as unknown.

The contribution of our work is three-fold. First, SOUL is the first approach developed to tackle the challenge of semantic segmentation on tropical forest ULS point clouds. SOUL demonstrates better wood point classification in complex tropical forest environments while exclusively utilizing point coordinates as input. Experiments show that SOUL exhibits promising generalization capabilities, achieving good performance even on data sets from other LiDAR devices, with a particular emphasis on overstory. Secondly, we propose a novel data preprocessing method, GVD, used to pre-partition data and address the difficult challenge of training neural networks from sparse point clouds in tropical forest environments.

Third, we mitigate the issue of imbalanced classes by proposing a new loss function, referred to as rebalanced loss function, which is easy to use and can work as a plug-and-play for various deep learning architectures. The data set (Bai et al. [15]) used in the article is already available in open access at https://zenodo.org/record/8398853 and our code is available at https://github.com/Na1an/phd_mission.

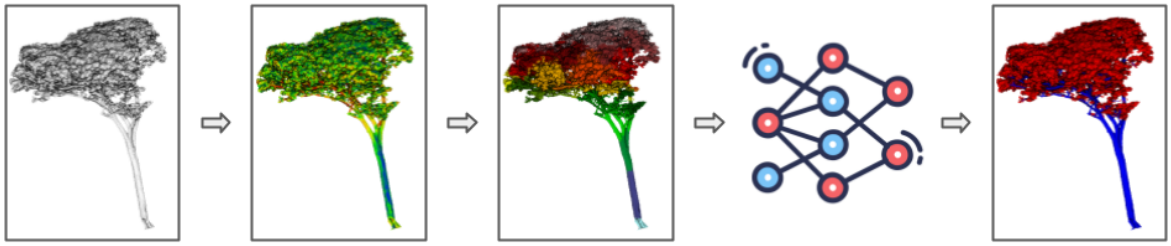



Figure 2: Overview of SOUL. (a) We use only the coordinates of raw LiDAR data as input. (b) Four geometric features linearity, sphericity, verticality, and PCA1 are calculated at three scales using eigenvalues, then standardized. (c) GVD proposes partitioned components and performs data normalization within these components. (d) Training deep neural network. (e) Finally, output are point-wise predictions.

# 2 Methodology

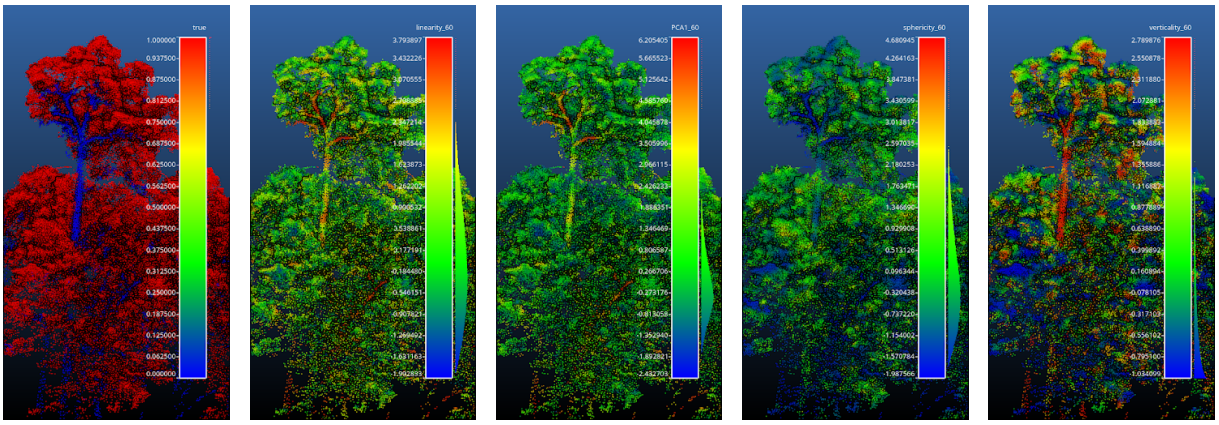
SOUL is based on PointNet++ Multi-Scale Grouping (MSG) [9] with some adaptations. The selection of PointNet++ is not only because of its demonstrated performance in similar tasks (Krisanski et al. [10]; Morel et al. [8]; Windrim & Bryson[11]), but also because of the lower GPU requirements (Choe et al. [16]) compared with transformer-based models developed in recent years, like the method proposed by Zhao et al. [17]. The main idea of SOUL lies in leveraging a geometric approach to extract preliminary features from the raw point cloud, these features are then combined with normalized coordinates into a deep neural network to obtain more abstract features in some high dimensional space [18]. We will introduce our method in detail as follows.

## 2.1 Data cleaning

Filter out returns below -20 dB, and eliminate noise and ground points.

## 2.2 Geometric feature computation

At this stage, we introduce four point-wise features: linearity, sphericity, verticality, and PCA1, which are computed at multiple scales of 0.3 m, 0.6 m, and 0.9 m in this task.



(a) True (b) Linearity (c) PCA1 (d) Sphericity (e) Verticality

Figure 3: Four point-wise features introduced at this stage.

## 2.3 Data pre-partitioning

The geodesic voxelization decomposition (GVD) algorithm is applied to partition the ULS data while preserving the topology of the point cloud. This approach enables the extraction of a set of representative training samples from raw forest data, while preserving the local geometry information in its entirety.

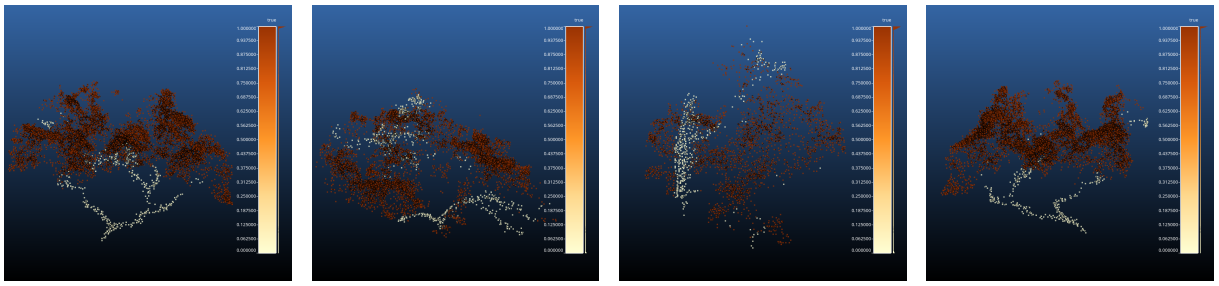


Figure 4: Component examples generated by the GVD algorithm.

## 2.4 Training neural network

Within the labeled ULS training data set, only 4.4% are wood points. The model is overwhelmed by the predominant features of leaves. Therefore, we propose a rebalanced loss $L_R$ that changes the ratio of data participating to 1:1 at the loss calculation stage by randomly selecting a number of leaf points equal to the number of wood points.

**Rebalanced loss**, denoted as $L_R$, is used to address class imbalance issue.

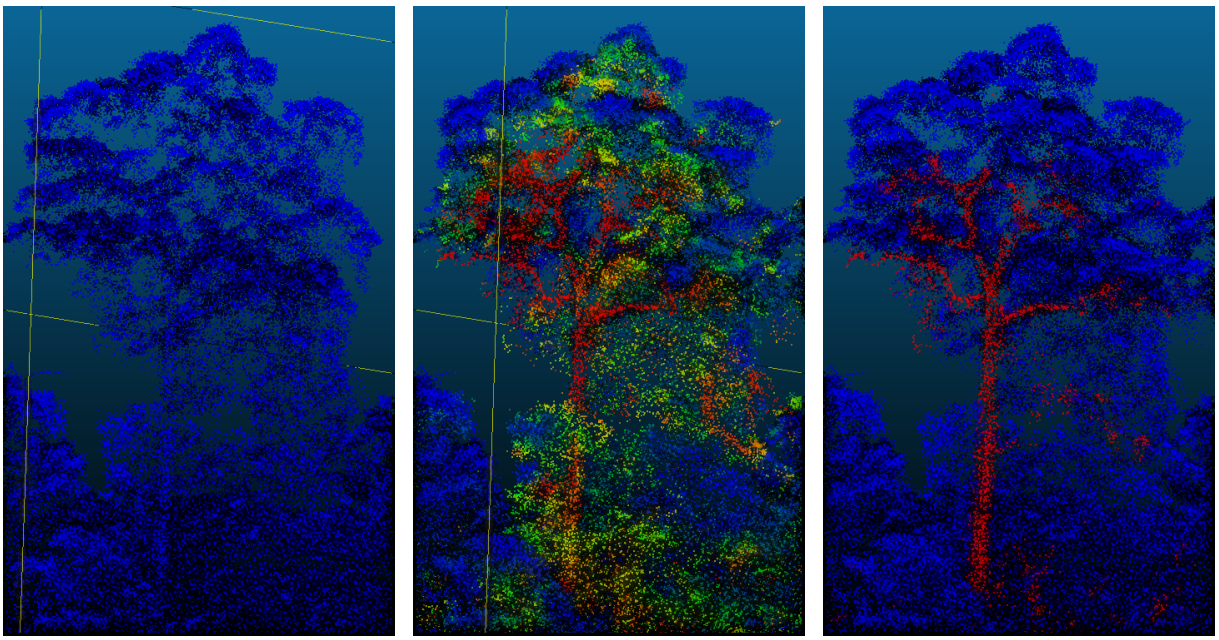
$$L_R(Y_{B_k}) = -\sum y_k \log(\hat{p}_k) + (1 - y_k) \log(1 - \hat{p}_k)\,,\ y_k \in (B'_{k,0} \cup B_{k,1}). \tag{1}$$

where $Y_{B_k}$ specifies the ground truth label of the batch $B_k$, $\hat{p} \in [0, 1]$ is the model's estimated probability for the label $y = 1$ and $B'_{k,0}$ is defined as

$$B'_{k,0} = \begin{cases} \text{downsampling}(B_{k,0}, |B_{k,1}|)\,, & \text{if } |B_{k,0}| \geq |B_{k,1}| \\ B_{k,0}\,, & \text{otherwise.} \end{cases} \tag{2}$$

# 3 Results

Comparatively to the prevailing methods employed for forest point clouds, our SOUL approach improves semantic segmentation on ULS forest data by large margins. The issue of class imbalance has been addressed.



(a) FSCT (b) SOUL (c) Ground Truth

Figure 5: Qualitative results on various LiDAR data from different sites.

# 4 Conclusions

We present SOUL, a novel approach for semantic segmentation in complex forest environments. It outperforms existing methods in the semantic segmentation of ULS tropical forest point clouds and demonstrates high performance metrics on labeled ULS data and generalization capability across various forest data sets. The proposed GVD method is introduced as a spatial split schema to provide refined training samples through pre-partition. One key aspect of SOUL is the use of the rebalanced loss function, which prevents drastic changes

in gradients and improves segmentation accuracy. While SOUL shows good performance for different forest types, it may struggle with significantly different trees without retraining. Future work can focus on improving the performance of SOUL on denser forest point clouds to broaden its applications.

# References


[1] Y. BAI, J.-B. Durand, G. L. Vincent, and F. Forbes, "Semantic segmentation of sparse irregular point clouds for leaf/wood discrimination," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[2] G. Vincent, C. Antin, M. Laurans, J. Heurtebize, S. Durrieu, C. Lavalley, and J. Dauzat, "Mapping plant area index of tropical evergreen forest by airborne laser scanning. A cross-validation study using LAI2200 optical sensor," *Remote Sensing of Environment*, vol. 198, pp. 254–266, 2017.

[3] A. Ullrich and M. Pfennigbauer, "Categorisation of full waveform data provided by laser scanning devices," in *Electro-Optical Remote Sensing, Photonic Technologies, and Applications V* (G. J. Bishop, J. D. Gonglewski, and G. W. Kamerman et al., eds.), vol. 8186, p. 818609, International Society for Optics and Photonics, SPIE, 2011.

[4] B. Brede, H. M. Bartholomeus, N. Barbier, F. Pimont, G. Vincent, and M. Herold, "Peering through the thicket: Effects of UAV LiDAR scanner settings and flight planning on canopy volume discovery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 114, p. 103056, 2022.

[5] G. Vincent, P. Verley, B. Brede, G. Delaitre, E. Maurent, J. Ball, I. Clocher, and N. Barbier, "Multi-sensor airborne lidar requires intercalibration for consistent estimation of light attenuation and plant area density," *Remote Sensing of Environment*, vol. 286, p. 113442, 2023.


| Methods | Accuracy | Recall | Precision | Specificity | G-mean | BA[1] |
|---|---|---|---|---|---|---|
| FSCT | 0.974 | 0.977 | 0.997 | 0.13 | 0.356 | 0.554 |
| FSCT + retrain | 0.977 | 1.0 | 0.977 | 0.01 | 0.1 | 0.505 |
| LeWos | 0.947 | 0.97 | 0.975 | 0.069 | 0.259 | 0.520 |
| LeWos (SoD[2] ) | 0.953 | 0.977 | 0.975 | 0.069 | 0.260 | 0.523 |
| SOUL (focal loss) | 0.942 | 0.958 | 0.982 | 0.395 | 0.615 | 0.677 |
| SOUL (rebalanced loss) | 0.826 | 0.884 | 0.99 | **0.631** | **0.744** | **0.757** |

[1] BA (Balanced Accuracy) $BA = \frac{1}{2}(Recall + Specificity)$.
[2] SoD (Significant of Difference).

Table 1: Comparison of different methods

[6] B. Wu, G. Zheng, and Y. Chen, "An Improved Convolution Neural Network-Based Model for Classifying Foliage and Woody Components from Terrestrial Laser Scanning Data," *Remote Sensing*, vol. 12, no. 6, 2020.

[7] D. Wang, S. Momo Takoudjou, and E. Casella, "LeWoS: A universal leaf-wood classification method to facilitate the 3D modelling of large tropical trees using terrestrial LiDAR," *Methods in Ecology and Evolution*, vol. 11, no. 3, pp. 376–389, 2020.

[8] J. Morel, A. Bac, and T. Kanai, "Segmentation of Unbalanced and In-Homogeneous Point Clouds and Its Application to 3D Scanned Trees," *Vis. Comput.*, vol. 36, p. 2419–2431, oct 2020.

[9] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Curran Associates, Inc., 2017.

[10] S. Krisanski, M. S. Taskhiri, S. Gonzalez Aracil, D. Herries, A. Muneri, M. B. Gurung, J. Montgomery, and P. Turner, "Forest Structural Complexity Tool—An Open Source, Fully-Automated Tool for Measuring Forest Point Clouds," *Remote Sensing*, vol. 13, no. 22, 2021.

[11] L. Windrim and M. Bryson, "Detection, Segmentation, and Model Fitting of Individual Tree Stems from Airborne Laser Scanning of Forests Using Deep Learning," *Remote Sensing*, vol. 12, no. 9, 2020.

[12] T. He, H. Huang, L. Yi, Y. Zhou, C. Wu, J. Wang, and S. Soatto, "GeoNet: Deep Geodesic Networks for Point Cloud Analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6881–6890, June 2019.

[13] M. Liu, X. Zhang, and H. Su, "Meshing Point Clouds with Predicted Intrinsic-Extrinsic Ratio Guidance," in *European Conference on Computer Vision (ECCV), Glasgow, UK* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 68–84, Springer International Publishing, August 2020.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[15] Y. Bai, G. Vincent, N. Barbier, and O. Martin-Ducup, "UVA laser scanning labelled las data over tropical moist forest classified as leaf or wood points," Oct 2023.

[16] J. Choe, C. Park, F. Rameau, J. Park, and I. S. Kweon, "PointMixer: MLP-Mixer For Point Cloud Understanding," in *European Conference on Computer Vision (ECCV)*, 2022.

[17] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

[18] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

# Smoothed Bootstrap et génération de données synthétiques pour la modélisation des extrêmes

Samuel Stocksieker [1,2] & Denys Pommeret [2] & Arthur Charpentier [3]

[1] *Université Claude Bernard Lyon 1, Laboratoire de Sciences Actuarielle et Financière (UR SAF), Lyon, 69007, France, samuel.stocksieker@univ-amu.fr*

[2] *Aix Marseille Univ, CNRS, I2M, Marseille, France, denys.pommeret@univ-amu.fr*

[3] *Université du Québec à Montréal, Canada, charpentier.arthur@uqam.ca*

**Résumé.** En apprentissage supervisé, il est assez fréquent de se retrouver confronté à des données présentant des distributions déséquilibrées. Cette situation entraîne souvent une difficulté d'apprentissage pour les algorithmes standards. La recherche et les solutions en matière d'apprentissage à partir de distributions déséquilibrées se sont principalement concentrées sur les tâches de classification. Malgré son importance, très peu de solutions existent pour la régression déséquilibrée (*Imbalanced Regression*). Dans cet article, nous proposons une procédure d'augmentation de données, nommée DENIS, basée sur des estimations à noyau de densité. Cette approche fournit une expression des densités conditionnelles des générateurs. Nous appliquons DENIS en régression déséquilibrée et proposons de le combiner à une nouveau type de générateur de type wild-boostrap pour simuler la variable cible, conditionnellement aux nouvelles données synthétiques. Nous évaluons les performances de l'algorithme DENIS dans des situations de régression déséquilibrée. Nous évaluons empiriquement et comparons notre approche et démontrons une amélioration significative par rapport aux techniques existantes.

**Abstract.** In supervised learning, it is quite common to encounter data with imbalanced distributions. This situation often leads to learning difficulties for standard algorithms. Research and solutions in imbalanced learning have mainly focused on classification tasks. Despite its importance, very few solutions exist for imbalanced regression. In this paper, we propose a data augmentation procedure, called DENIS, based on kernel density estimates. This approach provides an expression of the conditional densities of the generators. We apply DENIS in imbalanced regression by combining such generation procedures with a bootstrap resampling technique for the target values. We evaluate the performance of the DENIS algorithm in imbalanced regression situations. We empirically evaluate and compare our approach and demonstrate significant improvement over existing techniques.

**Keywords.** Smoothed Bootstrap, Kernel Density Estimate, Imbalanced, Synthetic Data

# 1 Introduction

Many real-world forecasting problems are based on predictive models in a supervised learning framework and the standard algorithms fail when the target variable is skewed. The learning from imbalanced data concerns many problems with numerous applications in different fields

[27], [21]. The major part of such works concerns imbalanced classification (see for instance [7] or [9] or [13] or [26] or [44]. As shown by [4], many solutions for dealing with imbalanced learning propose a pre-processing strategy especially the generation of new synthetic data. A large part of these existing methods consists of combining the well-known SMOTE algorithm [22].

In the literature, the imbalanced regression corresponds to *the correct prediction of rare extreme values of a continuous target variable* [22] but, contrary to the classification tasks, there is no level to quantify the imbalance and the labels are continuous. Unlike in a classification context, learning from an imbalanced dataset for regression tasks leads to two additional problems: i) the definition of the imbalanced phenomenon and ii) the identification of the observations that are considered as minority. Regression tasks over imbalanced data are not as well explored. Few works have addressed the problem despite the importance of this topic. The first and main works on this topic propose to binarize the problem with a relevant function and an associate threshold [40] in order to adapt some Imbalanced classification solutions [41], [5], [6], [31], [37], [8] for instance. This methodology presents the disadvantage of dividing the continuous distribution of the target variable into classes and therefore involves a loss of information. More recently, new other methods have emerged by using deep learning approaches such as [33], [16], or [23]. [45] proposes to use kernel density estimates to improve learning from imbalanced data with continuous targets.

This paper aims to propose a novel approach, which we shall call DENIS (for Data Enrichment for Numerical Imbalanced Situations) to deal with the imbalanced regression problem. The first step of DENIS consists of generating synthetic data for covariates based on kernel density estimators. This data augmentation procedure can be used independently to generate synthetic data (supervised or non-supervised). A second step of DENIS is concerned with the imbalanced regression where the generator model is combined with a wild-bootstrap procedure to generate target values given the synthetic covariates. The DENIS algorithm can be easily used in an imbalanced classification where the label of the target variable remains unchanged. The main contributions of our paper can be summarized as follows: i) Propose a global form to group some existing perturbation generators; ii) Deducing new synthetic data oversampling methods; iii) when combined with a wild-bootstrap the generator model provides a new method for dealing with imbalanced regression iv) we empirically compared our generalized algorithm and its variants with several state-of-the-art approaches and we obtained great performance on several datasets.

The paper is organized as follows. In Section 2 we give a general form of our data augmentation procedure corresponding to the first step of DENIS. We study two standard perturbation methods that are included in this approach. In Section 3 we develop the theory to obtain new generators and we will look more closely at the imbalanced regression. This is the second step of DENIS where we combine generators with a wild bootstrap to generate synthetic target values. Numerical results on several applications are presented in Section 5.

# 2 A New Kernel-Based Oversampling Approach

## 2.1 General Formulation

We consider a sequence of observations $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$, which are realizations of $n$ iid random variables $(\boldsymbol{X}, Y)$, where the target variable $Y$ is univariate and the covariate $\boldsymbol{X}$ is a $p$-dimensional random vector. The components of $\boldsymbol{X} = (X_1, \cdots, X_p)$ are supposed to be continuous or discrete and $Y$ is supposed to be either qualitative (classification) or quantitative (regression). Write $\tilde{\boldsymbol{x}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ the set of all observations. We propose a generalized oversampling procedure based on the form of the following weighted kernel density estimate:

$$g_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}}) = \sum_{i \in \mathcal{I}} \omega_i K_i(\boldsymbol{x}^*, \tilde{\boldsymbol{x}}), \tag{1}$$

where $(K_i)_{i\in\mathcal{I}}$ is a collection of kernels, $(\omega_i)_{i\in\mathcal{I}}$ is a sequence of positive weights with $\sum_{i\in\mathcal{I}} \omega_i = 1$, and $\mathcal{I}$ represents a subset of $\{1, 2, \cdots, n\}$. Here the index $^*$ stands for the synthetic data. In (1) we propose a general form for the conditional density for the synthetic data generators. The objective is to use the flexibility of the kernels to estimate the density of covariates in order to obtain synthetic data that reflects the distribution of the observations.

We propose to show that (1) generalizes perturbation-based synthetic data oversampling. We give an illustration with the basic algorithms ROSE and Gaussian Noise in Sections 2.2 where we demonstrate that these methods are particular cases of the generalized form (1), with corresponding parameters. In Section 3 we will show that some new methods can be deduced from the generic form (1) and we will compare some of them to current competitors in the imbalanced regression context.

The generators in (1) can be considered as smoothed bootstrap methods ([35], [24], [14]). Indeed, the smoothed bootstrap consists in drawing samples from kernel density estimators of the distribution. This bootstrap can be decomposed into two steps: first, a seed is randomly drawn and second, a random noise from the kernel density estimator is added to obtain a new sample. In the form (1), the first step is represented by the drawing weight $\omega_i$ and the second by the kernel $K_i(x)$. Convergence properties of smoothed bootstrap are given in [15] and [20]. As described by the authors, the smoothed bootstrap *provides better performances than classical bootstrap when a proper choice of smoothing parameters is used.* They proved the consistency of the smoothed bootstrap with the classical multivariate kernel estimator and more specifically the convergence in Mallow's metric. Other works have focused on the consistency of the multivariate kernel density estimate and proposed a relevant bandwidth matrix, for instance, [34], [32] and [19]. Note also that a kernel density estimator is a special case of mixture models with as many components as observations.

## 2.2 Rewriting Perturbation Approaches

We illustrate (1) by recovering two classical data augmentation procedures as follows:

- At each step of the ROSE algorithm (see [30]) the seed $S$ is selected randomly. Given $S$

synthetic data is generated with a multivariate density

$$g^{ROSE}_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}}, S=i) = K^{ROSE}_{H_n}(\boldsymbol{x}^* - \boldsymbol{x}_i) = \frac{1}{|H_n|^{1/2}} K(H_n^{-1/2}(\boldsymbol{x}^* - \boldsymbol{x}_i))),$$

where $K$ denotes the multivariate Gaussian kernel and $H_n = diag(h_1, ..., h_p)$ is the bandwidth matrix proposed by Bowman and Azzalini [2], with $h_q = (\frac{4}{(p+2)n})^{1/(p+4)}\widehat{\sigma}_q, q = 1, ..., p$. Finally, a synthetic random variable $\mathbf{X}^*$ is generated with the density

$$g^{ROSE}_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}}) = \frac{1}{n}\sum_{i=1}^{n} K^{ROSE}_{H_n}(\boldsymbol{x}^* - \boldsymbol{x}_i) = \sum_{i=1}^{n} \omega_i K^{ROSE}_{H_n}(\boldsymbol{x}^* - \boldsymbol{x}_i)).$$

- Similarly to ROSE, at each step of the Gaussian Noise algorithm (see [29]) a seed is selected and synthetic data is generated. Finally, the generating multivariate density has the form

$$g^{GN}_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}}) = \frac{1}{n}\sum_{i=1}^{n} K^{GN}_{H_n}(\boldsymbol{x}^* - \boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|H_n|^{1/2}} K(H_n^{-1/2}(\boldsymbol{x} - \boldsymbol{x}_i)),$$

where $H^{GN}_n = diag(h_1, ..., h_p)$, $h_q = \sigma_{noise}\widehat{\sigma}_q, q = 1, ..., p$.

Both cases are particular cases of (1) with $\omega_i = \frac{1}{n}$ and $K_i(\tilde{\boldsymbol{x}}, \boldsymbol{x}) = K_{H_n}(\boldsymbol{x} - \boldsymbol{x}_i)$, i.e. the same Gaussian kernel for all observations but with a different bandwidth matrix.

However, with these generators, the directions in the data space are randomly generated and so they can more explore the space. The distance between the new sample and the seed is also unbounded. However, the directions are randomly chosen and do not respect the correlation between the data and their support and the correlations between variables.

# 3 New Kernel-Based Methods

As the ROSE and GN techniques use a multivariate Gaussian kernel estimate with a diagonal bandwidth matrix, we can rewrite their associated generating density as follows:

$$g_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}}) = \sum_{i=1}^{n} \omega_i \prod_{j=1}^{p} K_{h_j}(x^*_j - x_{ij}) \tag{2}$$

with $K_{h_j}(u) = (2\pi)^{-1/2} h_j^{-1} e^{-\frac{1}{2{h_j}^2}u^2}$ the univariate gaussian kernel density estimator with smoothing parameter $h_j$. Such kernels are clearly not adapted for asymmetric, bounded or discrete variables. This remark is also true for the work of [45] which uses some symmetric kernels to improve the learning of imbalanced datasets. Another remark about this work is the division of the target variable support into $B$ groups that involve a loss of information.

To fix the drawback of the classical kernel we extend (2) by adapting (1) to the support of $\boldsymbol{x}$, considering some non-classical kernels (we refer to some works handling the kernel density

estimation for specific distributions inspired from [1], [36], [25], [12]). We suggest rewriting the form (1) as

$$g^{per}_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}}) = \sum_{i\in\mathcal{I}} \omega_i \prod_{j=1}^{p} K_{h_j}(x^*_j, x_{ij})$$

where $K_{h_j}(u,x)$ is a univariate kernel adapted to the nature of the $j$th variable and specifically defined on $x$ as follows:

- Gaussian kernel for a variable defined on $\mathbb{R}$ (classical kernel):

$$K_h(u,x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-x}{h})^2}.$$

- Binomial kernel for a discrete variable defined on $\mathbb{N}$:

$$K_h(u,x) = \frac{(x+1)!}{u!(x+1-u)!}\left(\frac{x+h}{x+1}\right)^u \left(\frac{1-h}{x+1}\right)^{x+1-u}.$$

- Gamma kernel for a positive asymmetric distribution defined on $[a,+\infty]$:

$$K_h(u,x) = \frac{u^{(x-a)/h}}{\Gamma(1+(x-a)/h)h^{1+(x-a)/h}} \exp\big(\frac{-u}{h}\big)\mathbb{1}_{[a,+\infty]}(u).$$

- Negative Gamma kernel for a negative asymmetric distribution defined on $[-\infty,b]$:

$$K_h(u,x) = \frac{u^{-(x-b)/h}}{\Gamma(1-(x-b)/h)h^{1-(x-b)/h}} \exp\big(\frac{-u}{h}\big)\mathbb{1}_{[-\infty,b]}(u).$$

- Beta kernel for a variable defined on $[0,1]$:

$$K_h(u,x) = \frac{u^{x/h}(1-u)^{(1-x)/h}}{\mathcal{B}(\frac{x}{h}+1, \frac{1-x}{h}+1)}\mathbb{1}_{[0,1]}(u).$$

- Truncated Gaussian kernel for a variable defined on $[a,b]$:

$$K_h(u,x) = \frac{\alpha}{h\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-x}{h})^2}\mathbb{1}_{[a,b]}(u), \alpha := \left(\int_a^b \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-x}{h})^2}\right)^{-1}$$

Note that if the Dirac kernel ($\mathbb{1}_{\boldsymbol{x}=\boldsymbol{x}_i}$) is used, we get the standard bootstrap: 1 includes also the simple oversampling. It is important to note that the DENIS algorithm uses an estimation of the smoothing parameter $h$ provided by some specific R-package dedicated to the density estimation (for instance, it uses the Silverman estimation for the Gaussian kernel). Their estimates are based on properties of univariate consistency. Another technique to deal with skewed or heavy-tailed distributions is to apply a transformation of the data in order to use classical kernel density estimation [10], [11] but it necessitates proposing a relevant transformation.

**Remark:** The use of a diagonal bandwidth matrix in (2) does not take into account the correlation between variables. To improve this issue, we could consider a full (symmetric positive definite) smoothing matrix. In that case, we would use a multivariate kernel density estimate considering the correlation between the variables which would be optimal for generating data. However, the estimation of this kind of matrix can be biaised because based on correlation matrix i.e. linear correlations. The thesis of Duong [17] and his associated R-package [18] proposed a Multivariate Gaussian Kernel but is a very time-expensive and offer a maximum of 6 dimensions. These works, despite their high quality, are very limited in practice since on the one hand datasets contain generally more than 6 variables, and on the other hand Gaussian Kernel estimators are inappropriate for mixed data.

# 4 DENIS as a Solution for Imbalanced Regression

Using the previous methods proposed we can generate synthetic covariates $X^*$. We then have to generate the target variable $Y$ given $X^*$ obtained from the first generator step of DENIS. In this sense we propose here to define the drawing weights $\omega_i$ as the inverse of the kernel density estimate for the target variable $Y$: the more isolated an observation is, the higher its drawing weight. The same idea is proposed in [38]. We also defined some safeguards to avoid getting some weights too high. Finally, the target variable is not generated in the same way as the covariate. Once the covariates are generated from our generator models, we propose to adapt a wild-bootstrap technique[1] [42] as follows: i) train a Random Forest on the initial sample; ii) predict target variable $\widehat{y}_i$; iii) draw uniformly a prediction error $\epsilon_k$ to generate a new $y_i^* := \widehat{y}_i + \epsilon_k v_i$ where $v_i$ is a random variable. We suggest an adaptation of this method to synthetic data in considering the impact of getting new covariate $y_i^* := \widehat{y}_i + |\epsilon_k| v_i \times sign(\widehat{y}_i - \widehat{y_i^*})$. This form is close to the Wild Bootstrap version with the Rademacher distribution. This is the second step of DENIS. (giving lower performances in our applications). The idea behind this proposition is to consider the prediction error and the impact of the synthetic covariate on the target variable. The choice of using a random forest is justified by its good predictive performance, its non-parametric nature, and the possibility of getting an error distribution for the same value of the target variable.

# 5 Application in Imbalanced Regression

Although the DENIS algorithm can be applied for the classification tasks, we focus on the imbalanced regression context because of the natural capacity of the form (1) to handle continuous variables. We test our approach on several real data set from a repository provided as a benchmark for imbalanced regression problems[2] and presented in [6].We compare our results to existing methods to deal with imbalanced regression from the *UBL* R-package ([3]): classical oversampling, SMOTE, Gaussian Noise, SMOGN, WERCS and ADASYN from the

[1]The kernel regression (Nadaraya-Watson estimator) was also tested but not selected because its high computation time and poor performance

[2]`https://paobranco.github.io/DataSets-IR/`

python-package *ImbalancedLearningRegression* ([43]). These techniques are used with their automatic relevance function and the same parameters as DENIS if any. To avoid sampling effects and obtain a distribution of prediction errors we ran 10 train-test datasets. In the same way, to avoid getting results dependent on some learning algorithms we use 10 models from the *autoML of the H2O R-package* [28] among the following algorithms: Distributed Random Forest, Extremely Randomized Trees, Generalized Linear Model with regularization, Gradient Boosting Model, Extreme Gradient Boosting and a Fully-connected multi-layer artificial neural network. It is also possible to use a clustering (Gaussian Mixture Model) in DENIS to apply a generation by cluster. Note that the ROSE algorithm did not exist for the imbalanced regression. We train, with the autoML, the following train dataset:

- Benchmark: UBL-Oversampling (*UBL-OS*), UBL-SMOTE for regression (*UBL-SMOTE*), UBL-Gaussian Noise for regression (*UBL-GN*), UBL-SMOGN for regression (*UBL-SMOGN*), UBL-WERCS (*UBL-WERCS*), IRL-ADASYN (*IRL-OS*)
- DENIS (step 1): Oversampling (*G-OS*), Gaussian Noise (*G-GN*), Gaussian Noise with GMM-clustering (*G-GNwCl*), ROSE (*G-ROSE*), ROSE with a GMM-clustering (*G-ROSEwCl*), Non classical Smoothed Boostrap with contraints on the distributions (*G-NCSB*), Classical Smoothed Boostrap (*G-CSB*).

Figure 1 presents RMSE gain (wrt the imbalanced dataset) and the median of the RMSE ranking. We can observe on these datasets that the DENIS algorithm empirically outperforms the state-of-the-art techniques, especially the Non-Classical Smoothed Bootstrap. The RMSE-rank represents the ranking of approaches according to the RMSE for a run: rank 1 corresponds to the training dataset that offers the smallest RMSE on the test sample. Beyond performance in prediction, non-classical kernels allow for generating more coherent synthetic data by preserving the domains of definition of the variables.

| RMSE gain | NO2 | cpuSm | Boston | Bank8FM | Abalone |
|---|---|---|---|---|---|
| UBL-OS | -3% | 2% | -7% | 63% | 0% |
| UBL-SMOTE | -10% | -2% | -12% | 27% | -4% |
| UBL-GN | -8% | -5% | -7% | 57% | -3% |
| UBL-SMOGN | -10% | -3% | -10% | 29% | -3% |
| UBL-WERCS | -4% | 3% | -1% | 57% | -1% |
| IRL-ADASYN | 10% | 22% | -1% | 69% | NA |
| G-OS | -1% | -2% | -3% | 57% | -3% |
| G-GN | -11% | -18% | -21% | 0% | -9% |
| G-GNwCl | -10% | -18% | -14% | 13% | -6% |
| G-ROSE | -9% | -17% | -17% | -6% | -9% |
| G-ROSEwCl | -9% | -19% | -21% | 0% | -8% |
| G-NCSB | -9% | -23% | -23% | -6% | -9% |
| G-CSB | -12% | -20% | -17% | 0% | -9% |

(a) RMSE gain

| RMSE rank | NO2 | cpuSm | Boston | Bank8FM | Abalone |
|---|---|---|---|---|---|
| Imb | 16,0 | 13,0 | 16,5 | 7,0 | 16,5 |
| UBL-OS | 14,0 | 14,5 | 13,0 | 17,0 | 16,5 |
| UBL-SMOTE | 10,0 | 11,5 | 11,0 | 11,0 | 12,0 |
| UBL-GN | 10,5 | 9,5 | 11,5 | 15,5 | 13,0 |
| UBL-SMOGN | 8,0 | 12,0 | 11,5 | 12,0 | 12,0 |
| UBL-WERCS | 14,0 | 16,0 | 15,5 | 16,5 | 15,0 |
| IRL-ADASYN | 18,0 | 18,0 | 16,0 | 18,0 | NA |
| G-OS | 16,0 | 14,0 | 14,5 | 15,5 | 13,5 |
| G-GN | 6,5 | 5,5 | 2,5 | 6,0 | 6,5 |
| G-GNwCl | 8,5 | 5,0 | 6,5 | 8,0 | 9,5 |
| G-ROSE | 8,5 | 4,0 | 4,0 | 4,5 | 7,0 |
| G-ROSEwCl | 10,0 | 4,0 | 3,0 | 8,0 | 8,0 |
| G-NCSB | 8,0 | 2,0 | 4,0 | 5,0 | 5,5 |
| G-CSB | 6,5 | 3,5 | 6,0 | 5,0 | 6,0 |

(b) Median of the RMSE-rank

Figure 1: RMSE-gain and median of the RMSE-rank on the Imbalanced Regression Datasets

We can see on these several applications, with several runs, several learning algorithms, and several performance metrics that the DENIS approach seems relevant to deal with imbalanced regression.

# 6 Discussion and Perspectives

DENIS is a new approach that offers a general form for the generator which is based both on the theoretical foundations of kernel estimators and classical smoothed bootstrap techniques. It provides a general expression for the conditional density of the generator. The use of well-chosen kernels makes it possible to take into account the nature of the covariates: continuous, discrete, totally or partially bounded. Numerical applications in imbalanced regression models demonstrate that DENIS and its variants are very competitive.

The weights $\omega_i$ offer a large flexibility to use the method. For instance, it is possible to handle classification tasks by conditioning with the minority class. We could deal with multi-class classification too. It is also possible to combine another weighting, proposed in the literature, to focus on specific samples in the synthetic data generation with a kernel approach to perform the methodology.

As a perspective, a natural extension of this work is to automate the choice of the kernel estimators, the weights, as well as some parameters according to the data. It will be possible to define a kernel according to the neighborhood in the same dataset. We also could define $\omega_i$ in order to generate to a target distribution as done in [39]. Finally, by nature, the perturbation approaches with a diagonal bandwidth matrix do not consider the correlation between covariates. The interpolation approaches consider it but the generation is limited to the segments. Another direction for further investigations would be to better consider the correlations between variables and be able to handle mixed data.

# References


[1] Taoufik Bouezmarni and Jeroen VK Rombouts. Nonparametric density estimation for multivariate bounded data. *Journal of Statistical Planning and Inference*, 140(1):139–152, 2010.

[2] Adrian W. Bowman and Adelchi Azzalini. Applied smoothing techniques for data analysis : the kernel approach with s-plus illustrations. *Journal of the American Statistical Association*, 94:982, 1999.

[3] Paula Branco, Rita P Ribeiro, and Luis Torgo. Ubl: an r package for utility-based learning. *arXiv preprint arXiv:1604.08079*, 2016.

[4] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, 49(2):1–50, 2016.

[5] Paula Branco, Luís Torgo, and Rita P Ribeiro. Smogn: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, pages 36–50. PMLR, 2017.

[6] Paula Branco, Luis Torgo, and Rita P Ribeiro. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343:76–99, 2019.

[7] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[8] Luís Camacho, Georgios Douzas, and Fernando Bacao. Geometric smote for regression. *Expert Systems with Applications*, page 116387, 2022.

[9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[10] Arthur Charpentier and Emmanuel Flachaire. Log-transform kernel density estimation of income distribution. *L'Actualité économique*, 91(1):141–159, 2015.

[11] Arthur Charpentier and Abder Oulidi. Beta kernel quantile estimators of heavy-tailed loss distributions. *Statistics and computing*, 20(1):35–55, 2010.

[12] Song Xi Chen. Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52:471–480, 2000.

[13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Classbalanced loss based on effective number of samples. *CVPR*, 2019.

[14] Daniela De Angelis and G Alastair Young. Smoothing the bootstrap. *International Statistical Review/Revue Internationale de Statistique*, pages 45–56, 1992.

[15] Daniele De Martini and Fabio Rapallo. On multivariate smoothed bootstrap consistency. *Journal of statistical planning and inference*, 138(6):1828–1835, 2008.

[16] Yifei Ding, Minping Jia, Jichao Zhuang, and Peng Ding. Deep imbalanced regression using cost-sensitive learning and deep feature transfer for bearing remaining useful life estimation. *Applied Soft Computing*, 127:109271, 2022.

[17] Tarn Duong. *Bandwidth selectors for multivariate kernel density estimation.* University of Western Australia Perth, 2004.

[18] Tarn Duong. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of statistical software*, 21:1–16, 2007.

[19] Tarn Duong and Martin L Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.

[20] M Falk and R-D Reiss. Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. *The Annals of Probability*, pages 362–371, 1989.

[21] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[22] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[23] Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. *arXiv preprint arXiv:2205.15236*, 2022.

[24] Peter Hall, Thomas J DiCiccio, and Joseph P Romano. On smoothing and the bootstrap. *The Annals of Statistics*, pages 692–704, 1989.

[25] Tristen Hayfield and Jeffrey S Racine. Nonparametric econometrics: The np package. *Journal of statistical software*, 27:1–32, 2008.

[26] C. Huang, Y. Li, C. Change Loy, and X. Tang. Learning deep representation for imbalanced classification. *CVPR*, 2016.

[27] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[28] Erin LeDell and Sebastien Poirier. H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*, July 2020.

[29] Lee and Sauchi. Noisy replication in skewed binary classification. *Computational Statistics and Data Analysis*, 34(2):165–191, 2000.

[30] Menardi and Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, 2014.

[31] Rita P Ribeiro and Nuno Moniz. Imbalanced regression and extreme value prediction. *Machine Learning*, 109:1803–1835, 2020.

[32] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[33] Snigdha Sen, Krishna Pratap Singh, and Pavan Chakraborty. Dealing with imbalanced regression problem for large dataset using scalable artificial neural network. *New Astronomy*, 99:101959, 2023.

[34] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[35] BW Silverman and GA Young. The bootstrap: to smooth or not to smooth? *Biometrika*, 74(3):469–479, 1987.

[36] Sobom Matthieu Someé. *Estimations non paramétriques par noyaux associés multivariés et applications*. PhD thesis, Université de Franche-Comté, 2015.

[37] Xin Yue Song, Nam Dao, and Paula Branco. Distsmogn: Distributed smogn for imbalanced regression problems. In *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 38–52. PMLR, 2022.

[38] Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110:2187–2211, 2021.

[39] Samuel Stocksieker, Denys Pommeret, and Arthur Charpentier. Data augmentation for imbalanced regression. In *International Conference on Artificial Intelligence and Statistics*, pages 7774–7799. PMLR, 2023.

[40] Luis Torgo and Rita Ribeiro. Utility-based regression. In *PKDD*, volume 7, pages 597–604. Springer, 2007.

[41] Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. Smote for regression. In *Portuguese conference on artificial intelligence*, pages 378–389. Springer, 2013.

[42] C. F. J. Wu. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261 – 1295, 1986.

[43] Wenglei Wu, Nicholas Kunz, and Paula Branco. Imbalancedlearningregression-a python package to tackle the imbalanced regression problem. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part VI*, pages 645–648. Springer, 2023.

[44] Y. Yang and Z. Xu. Rethinking the value of labels for improving class-imbalanced learning. *NeurIPS*, 2020.

[45] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, pages 11842–11851. PMLR, 2021.

# Inférence de probabilités prédictives à l'aide des forêts aléatoires dans le contexte de classification déséquilibrée

Clément Dombry [2], Moria Mayala [*1], Charles Tillier [3] & Olivier Wintenberger [1]

[1] *Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université*
[2] *Laboratoire de Mathématiques de Besançon – Université de Franche-Comté*
[3] *Laboratoire de Mathématiques – Université de Versailles Saint-Quentin-en-Yvelines*

**Résumé.** Les données déséquilibrées dans les taches de classification sont aujourd'hui identifiées comme un problème majeure en apprentissage automatique. Une de ces raisons est que les algorithmes traditionnels de machine learning peuvent être mis en difficulté dans ce cadre, notamment pour détecter la classe minoritaire. Dans ce travail, nous faisons de l'inférence des probabilités prédictives reposant sur les modèles simplifiés en particulier des forêts purement aléatoires infinies (IPRF) en vue de relever les défis associés à la prédiction d'événements rares. Nous établissons notamment un théorème central limite pour cet estimateur IPRF sous certaines hypothèses de régularité sur la fonction de régression. Cependant, IPRF hérite un biais asymptotique inhérent à l'asymetrie de la distribution de classes. Nous proposons une procédure de type échantillonnage préférentielle dérivant des odd-ratio afin de réduire le biais asymptotique de IPRF. Une courte étude de simulation illustre les performances de la méthode proposée.

**Mots-clés.** Classification binaire, données déséquilibrées, forêts purement aléatoires infinies.

**Abstract.** Imbalanced data in classification tasks has been identified as an top problem in machine learning. One of the reasons for this is that traditional machine learning algorithms can get into trouble in this context, particularly when it comes to detecting the minority class. In this work, we perform predictive probability inference based on simplified models (algorithms), in particular Infinite purely random forests (IPRFs), to address the challenges associated with predicting rare events. In particular, we establish a central limit theorem for this IPRF estimator under certain regularity assumptions on the regression function. However, IPRF inherits an asymptotic bias due to the asymmetry of the class distribution. We propose a preferential sampling procedure derived from the odd-ratio in order to reduce the asymptotic bias of IPRF. A short simulation study illustrates the performance of the proposed method

**Keywords.** binary classification, class imbalance, infinite random forests.

# 1 Contexte

Il est connu que la prédiction sur les données déséquilibrées répresente une difficulté majeure en apprentissage automatique, en particulier dans les modèles de classification, qu'il s'agisse de classification binaire (ex :détecter une maladie, détection d'images, intrusion réseau, détection de fraude ) ou de classification multi-classes (ex : prédire le modèle de voiture acheté), pour plus de détailles une bonne revue a été proposée par Jason Brownlee (2020). Ainsi, proposer une procedure éfficace d'estimation dans ce cadre est un enjeu d'intérêt. Nous présentons ici le cas binaire, qui est plus simple à appréhender et peut ensuite facilement se généraliser au multi-classes.

Le problème de base est le suivant. On considère $(X_i, Y_i)_{1\leq i\leq n}$ des observations i.i.d de même loi que le couple $(X, Y) \in \mathcal{X}\times\{0, 1\}$, $\mathcal{X} \subset \mathbb{R}^d$. La loi jointe de $(X, Y)$ est complètement déterminer par la loi marginale de $X$ et la fonction de régression $\mu$ est définie par

$$\mu(\mathbf{x}) := \mathbb{P}(Y = 1|X = \mathbf{x}), \qquad \mathbf{x} \in \mathcal{X}. \tag{1}$$

Notre objectif est d'estimer $\mu$ lorsque les classes $\{0, 1\}$ ne sont pas représentées de manière égale. Pour quantifier le degré de *déséquilibre*, nous utilisons la notion de ratio de déséquilibre ( Imbalance ratio ) définie comme suit

$$IR = n_0/n_1$$

où $n_0$ et $n_1$ désignent respectivement le cardinal des échantillons majoritaire et minoritaire, respectivement. Il ya un déséquilibre dans les données lorsque $IR > 1$ . Dès lors que la valeur $IR \approx 1$ correspondant à des données parfaitement équilibrées (cf. Robert O'Brien et al (2019)). Sans perte de généralité, nous supposons que la classe minoritaire correspond à l'etiquette 1.

## 1.1 Cadre statistique

On considère deux (2) modèles: le premier modèle est définit sous les mêmes hypothèses que celle de eq. (1). Il s'agit du modèle original d'intérêt, pour lequel nous aimerions prédire avec précision la classe d'appartenance des observations qui sont supposés être déséquilibrées. Ainsi, si on note

$$p := \mathbb{P}(Y = 1) = \int_{\mathcal{X}} \mu(\mathbf{x}) d\mathbb{P}_X(\mathbf{x}).$$

la probabilité d'observer la classe 1, on suppose donc que $p < q := \mathbb{P}(Y = 0)$.

Afin de gérer ce déséquilibre, on considère que l'on a également accès à des données provenant d'un deuxième modèle. Dans ce dernier, on suppose que les observations ne sont pas déséquibrées c'est à dire

$$p^* := \mathbb{P}(Y^* = 1) = \int_{\mathcal{X}} \mu^*(\mathbf{x}) d\mathbb{P}_{X^*}(\mathbf{x}),$$

avec $\mu^*(\mathbf{x}) = \mathbb{P}(Y^* = 1|X^* = \mathbf{x})$, pour $\mathbf{x} \in \mathcal{X}$, la fonction de régression et on a $p^* = q^*$. Nous supposons en outre la condition suivante :

- La probabilité de classe $p^* \in (0,1)$ satisfait $p^* > p$. Un choix pratique courant est $p^* \approx 0,5$.

A partir d'observations provenant de ces deux modèle, notre stratégie consiste à sous-échantillonner la classe majoritaire dans le premier modèle dans le but d'équlibrer les données.

.

## 1.2 Méthode: IPRF avec échantillonnage préférentiel

Introduit par Leo Breiman (2001) les forêts aléatoires, constituent un algorithme d'apprentissage très populaire qui offre d' excellente performances, et une grande flexibilité dans sa capacité à gérer tous les types de données. Ainsi dans le contexte des données déséquilibrées ( imbalanced data), Leo Breiman et al (2004) proposent une nouvelle variante de forêts aléatoires, simplifiées appelée forêts aléatoires équilibrées (balanced random forests, PRF) qui traite à tous points de vue le probleme d'imbalanced data en sous-échantillonnant la classe majoritaire afin d'améliorer les performances de classification par rapport à la classe minoritaire. Sans perte de généralité, les forêts purement aléatoires infinies (IPRF) interviennent dans le même contexte en vue de relever les défis associés à la prédiction d'événements rares. Notre approche s'appuit sur les travaux antérieurs de Wager et Athey (2018) et de Peng et al. (2022) qui ont établi la normalité asymptotique de variantes infinies similaires de différents estimateurs à l'aide des outils de U-statistique.

Notre objectif est de construire un classifier $\widehat{\mu}_{IS}(\mathbf{x})$ de $\mu(\mathbf{x})$. Etant donné, $p$ et $p^*$ les proportions de la classe 1 dans les modèles original et biaisé(second modèle) . A partir des ratio de chances on établit une connexion entre $\mu^*(\mathbf{x})$ et $\mu(\mathbf{x})$ donnée par

$$\frac{\mu(\mathbf{x})/\left(1-\mu(\mathbf{x})\right)}{p/(1-p)} = \frac{\mu^*(\mathbf{x})/\left(1-\mu^*(\mathbf{x})\right)}{p^*/(1-p^*)}\,, \qquad \mathbf{x} \in \mathcal{X}\,. \tag{2}$$

Par ailleurs, on estime les proportions de la classe 1 dans les deux modèles.

**Hypothès 1** *Soient les tailles des sous-échantillons dans la classe 0 et la classe 1 $s_0$ et $s_1$ dans le modèle biaisé et $n_0$ et $n_1$ les tailles d'échantillons de la classe 0 et la classe 1 dans le modèle original*

$$\widehat{p}^* := \frac{s_1}{s_1+s_0} \longrightarrow p^* > 0\,, \qquad n \to \infty\,.$$

*Remarquons que une valeur de $p^* = 0.5$ peut correspondre au cas des données équilibrées. De manière analogue la loi des grands nombre fournie l'estimation suivante de $p$*

$$\widehat{p} := \frac{n_1}{n_0+n_1} \to p\,, \qquad n \to \infty\,.$$

Par échantillonnage préferentiel, nous obtenons cet estimateur

$$\widehat{\mu}_{IS}(\mathbf{x}) := \frac{n_1 s_0 \widehat{\mu}^*(\mathbf{x})}{n_0 s_1\left(1-\widehat{\mu}^*(\mathbf{x})\right) + n_1 s_0 \widehat{\mu}^*(\mathbf{x})}\,, \qquad \mathbf{x} \in \mathcal{X}\,, \tag{3}$$

**Théorème 1** *Soit* $\widehat{\mu}_{IS}(\mathbf{x})$ *un estimateur par échantillonnage d'importance comme défini dans* (3)*. Sous l' Hypothèse 1 , nous obtenons le théorème limite central suivant*

$$\sqrt{np_n}\left(\frac{\widehat{\mu}_{IS}(\mathbf{x})}{\mu(\mathbf{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{1-\mu(\mathbf{x})}{\mu(\mathbf{x})}\right),\ \textit{quand } n \to \infty. \tag{4}$$

*Où* $p_n = \mathbb{P}(X \in L_b(\mathbf{x})) \leq \mathbb{E}\Big[\mathrm{Diam}(L_b(\mathbf{x}))^d\Big]$*, si* $X \sim \mathcal{U}([0,1]^d)/\mathcal{X} = [0,1]^d$*.*

## 1.3 Etude numérique

Dans cette section, nous donnons à l'aide de données synthétiques, une illustration numérique de nos résultats théoriques. L'objectif est de montrer que notre procédure d'inférence peut être mise en œuvre et de confirmer que les propriétés asymptotiques dérivées sont empiriquement pertinentes. Considérons les observations $(X_i, Y_i)_{1\leq i\leq n}$ i.i.d, où les covariables et la variable cible sont définies comme suit

- $X_i \sim \mathcal{U}([0,1]^2)$, et $\ Y_i \in \{0,1\}$
- $n = \{100 \ldots 1000\}$ et 30 répétitions Monte-Carlo
- Utilisation du package `RandomUniformForest` ( Saip Ciss,(2014)).

Nous évaluons la performance de $\widehat{\mu}_{IS} := g(\text{PRF-balanced})$ en terme de l'erreur quadratique moyenne (MSE). Cette métrique emblématique peut s'écrire comme une somme de la variance du modèle, du biais du modèle et de l'incertitude irréductible (compromis biais-variance). Par soucis de clarté, notons PRF $:= \widehat{\mu}^*$.

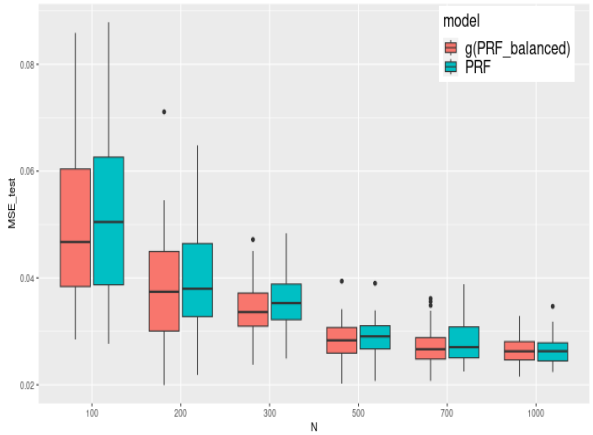



Figure 1: MSE de l'estimateur par échantillonnage d'importance $\widehat{\mu}_{IS}$ et de l'estimateur de forêts aléatoires équilibrées $\widehat{\mu}^*$ pour , $d = 2$, toutes les forêts ont $B = 500$ arbres.

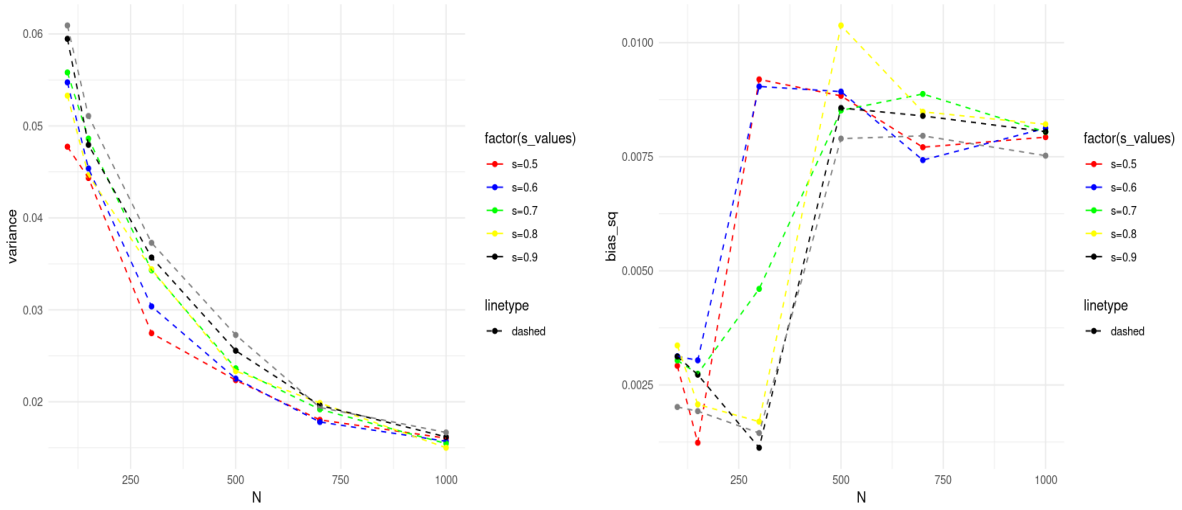



Figure 2: Variance (à gauche), Biais (à droite) de l'estimateur par échantillonnage d'importance $\widehat{\mu}_{IS}$ pour les différentes valeurs de $s \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\} * n$ , $d = 2$, toutes les forêts ont $B = 500$ arbres.

# Remerciements

Nous remercions Saip Ciss pour l'amélioration substantielle des hyperparamètres dans le package `randomUniformForest`. Nous remercions également Francesco Bonacina pour toute l'aide qu'il nous a apportée dans la réalisation de ces simulations. En outre, nous aimerions remercier plus particulièrement les membres du projet ANR T-REX pour le financement de nos multiples ateliers à Vienne et ailleurs au cours desquels des discussions fructueuses ont permis d'améliorer grandement ce document. Nous remercions également le "International Emerging Actions 2022" et le "Centre National de la Recherche Scientifique (CNRS)" pour le support de voyage de Moria Mayala et Charles Tillier à Vienne.

# Bibliographie

Leo Breiman.(2001), Random forests.,*Machine learning*, 45(1):5–32.

Chao Chen, Andy Liaw, Leo Breiman, et al.(2004), Using random forest to learn imbalanced data.,*University of California, Berkeley*, 10(1-12):24.

Jason Brownlee. (2020), *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning.*, Machine Learning Mastery.

Robert O'Brien and Hemant Ishwaran.(2019), A random forests quantile classifier for class imbalanced data.,*Pattern recognition*, 90:232–249.

Saïp Ciss. (2014), *Forêts uniformément aléatoires et détection des irrégularités aux cotisations sociales.*, PhD thesis, Paris 10.

Stefan Wager and Susan Athey. (2018), Estimation and inference of heterogeneous treatment effects using random forests., *Journal of the American Statistical Association* , 113 (523):1228–1242.

Sylvain Arlot and Robin Genuer.(2014), Analysis of purely random forests bias.,*arXiv preprint* , arXiv:1407.3939.

Wei Peng, Tim Coleman, and Lucas Mentch. (2022), ates of convergence for random forests via generalized u-statistics, *Electronic Journal of Statistics*, 16(1):232–292.

# Données de comptage

# PLN-tree: un modèle latent d'inférence de réseaux sur des arbres d'abondance au sein du microbiote

Alexandre Chaussard [1] & Anna Bonnet [1] & Sylvain Le Corff [1] & Harry Sokol [2]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, France, chaussard@lpsm.paris*

[2] *CHU Saint-Antoine Assistance publique Hôpitaux de Paris (AP-HP), Sorbonne Université, France*

**Résumé.** Le microbiote intestinal est un écosystème complexe composé principalement de bactéries qui interagissent entre elles et avec leur environnement. Si la composition du microbiote s'est avérée être un biomarqueur pertinent pour plusieurs maladies, la structure spécifique du microbiote intestinal a été peu prise en compte à ce jour. En effet, la composition microbienne est décrite par des données discrètes et parcimonieuses qui s'organisent selon une structure d'arbre taxonomique. Plus précisément, chaque bactérie appartient à plusieurs groupes qui sont hiérarchiquement ordonnés du plus précis (espèce) au moins précis (domaine). Bien que cette information taxonomique soit connue, on ignore encore son rôle dans l'impact d'une bactérie sur son hôte. Des travaux récents (Chiquet et al. 2019) ont proposé un cadre permettant de modéliser des données de comptage parcimonieuses pour étudier des interactions entre espèces, mais dans lequel l'impact de la structure taxonomique n'a pas encore été étudié. Dans ce travail, nous visons donc à étendre les travaux existants afin d'établir un cadre et un modèle dédié aux données de comptage s'appuyant sur une structure taxonomique. Notre objectif est de proposer des méthodes interprétables pour expliquer l'interaction complexe entre les espèces bactériennes, avec une application dans le contexte des maladies inflammatoires de l'intestin.

L'approche proposée, basée sur le modèle Poisson log-normal (PLN), tient compte des dépendances markoviennes pour inclure la structure de l'arbre taxonomique reliant les bactéries tout en permettant aux taxons de différentes branches de s'influencer mutuellement. Cela permet non seulement d'étendre le champ d'application de la modélisation des interactions, mais aussi de s'aligner sur la complexité et la diversité inhérentes aux communautés microbiennes. En outre, nous présentons une nouvelle approche variationnelle qui incorpore la structure markovienne de la distribution *a posteriori* afin d'améliorer la précision de l'estimation variationnelle.

Nous avons ensuite appliqué notre méthode aux données issues d'une cohorte de patients diagnostiqués avec la maladie de Crohn. Le modèle est systématiquement comparé à des modèles alternatifs, ce qui permet une analyse complète de sa capacité à détecter des interactions microbiennes complexes associées à des conditions pathologiques.

**Keywords.** modèles latents, PLN, microbiote, données compositionnelles, arbre

**Abstract.** The gut microbiota is a complex ecosystem composed mostly of bacteria interacting with each other and their environment. If the composition of the microbiota has

proven to be a relevant biomarker for several diseases, the specific structure of gut microbiota has poorly been taken into account. Indeed, the microbial composition is described by discrete and sparse data, which also have a taxonomic structure. More precisely, each bacterium belongs to several groups that are hierarchically ordered from more precise (species) to less precise (domain). Although this taxonomic information is known, it remains unclear how it is related to the impact of a bacterium on its host. While some recent works (Chiquet et al. 2019) proposed a framework that accounts for the sparse and discrete nature of the microbiota, the impact of the taxonomic structure has not been investigated yet. In this work, we aim at introducing a new framework and dedicated algorithms that account for the sparse taxonomic abundance nature of the microbiota data. Our focus is to propose interpretable methods to unveil the complex interplay among bacterial species, with an application in the context of inflammatory bowel diseases.

The proposed approach, based on Poisson Log-Normal models, accounts for Markov dependencies to include the taxonomic tree structure linking the bacteria while enabling taxa of different branches to influence each other. This not only extends the scope of interaction modeling but also aligns with the inherent complexity and diversity of microbial communities. Additionally, we present a novel variational approach that incorporates the Markovian structure of the posterior distribution to enhance the precision of the variational estimation. Such variational families allow us to obtain theoretical guarantees on latent state estimation and data reconstruction.

To assess real-world applications, we conduct an empirical investigation on a cohort of patients diagnosed with Crohn's disease. The model is systematically benchmarked against state-of-the-art algorithms, providing a comprehensive analysis of its efficiency in capturing complex microbial interactions associated with pathological conditions.

**Keywords.** latent models, PLN, microbiota, count data, tree

# 1 Introduction

Le microbiote intestinal humain est un écosystème complexe de $10^{13}$ micro-organismes résidant dans le tractus gastro-intestinal et qui joue un rôle clé dans la santé de l'hôte. Parmi les micro-organismes qui le composent, les bactéries constituent une part majoritaire, exerçant des effets profonds sur la physiologie, le métabolisme et les fonctions immunitaires de l'hôte. La compréhension des relations complexes au sein du microbiome intestinal est une étape cruciale vers la compréhension de son impact sur la santé humaine et la prédiction des résultats des traitements [Julien et al., 2022]. Dans cette optique, les progrès des technologies de séquençage à haut débit ont permis de caractériser les communautés microbiennes.

Parmi les différentes représentations du microbiote intestinal, nous nous concentrons sur les échantillons de taxa-abondance, qui permettent de représenter la composition d'un microbiome en combinant les données de comptage et les informations taxonomiques. Étant donné que la taxonomie fournit des informations sur les relations génétiques entre les espèces, cette structure d'arbre peut apporter des informations précieuses à la modélisation du nombre d'espèces. Cependant, les échantillons d'abondance taxonomique sont sujets à une grande variabilité inter-individuelle, ce qui complexifie le problème de la modélisation et de l'inférence. Outre la tâche de modélisation, nous visons à fournir un outil interprétable pour la recherche médicale en mettant l'accent sur les interactions entre espèces qui pourraient mieux expliquer le rôle des troubles du microbiome. Dans ce contexte, nous présentons un nouveau modèle pour les données de comptage associées à une taxonomie, PLNtree, conçu pour estimer les interactions microbiennes tout en modélisant les échantillons de taxa-abondance.

Le modèle PLNtree s'appuie sur le cadre Poisson Log-Normal (PLN), développé par [Aitchison and Ho, 1989] puis largement étudié par [Chiquet et al., 2021, Chiquet et al., 2019], qui modélise les données de comptage et les interactions entre espèces au moyen d'une variable latente gaussienne. Notre contribution consiste à étendre le modèle PLN et les méthodes d'inférence associées pour tenir compte de la structure arborescente inhérente à la taxonomie microbienne. En outre, pour garantir la flexibilité et l'adaptabilité, le modèle PLNtree incorpore une chaîne de Markov latente indépendante de la structure taxonomique qui capture la dynamique et les interactions sous-jacentes au sein de la communauté microbienne. Cette caractéristique permet d'explorer les relations cachées entre tous les taxons d'un niveau. Simultanément, les comptes observés suivent une chaîne de Markov contrainte qui adhère à la structure hiérarchique de la taxonomie, préservant ainsi l'intégrité de la composition des données de taxa-abondance.

De plus, comme l'apprentissage du modèle PLNtree nécessite l'utilisation d'un proxy variationnel, nous fournissons une approximation variationnelle "backward" amortie produisant des résultats affinés face à l'approche "mean-field".

Afin de valider l'efficacité du modèle PLNtree, nous présentons des résultats expérimentaux basés sur une cohorte de patients diagnostiqués avec la maladie de Crohn. Nous comparons le modèle PLNtree à des concurrents sur la tâche de génération en utilisant le critère de diversité alpha.

# 2 Cadre mathématique

## 2.1 Modèle Poisson Log-Normal model

La recherche d'interactions entre entités peut être formulée comme un problème d'inférence de réseau, pour lequel les méthodes canoniques sont basées sur les Modèles Graphiques Gaussiens [Altenbuchinger et al., 2020]. Toutefois, l'hypothèse de normalité ne s'applique pas aux données d'abondance. Des approches statistiques ont donc été proposées pour analyser les données d'abondance [Inouye et al., 2017]. En particulier, le modèle Poisson Log-Normal [Aitchison and Ho, 1989] a conduit à de nombreux développements théoriques et méthodologiques [Chiquet et al., 2019] dans l'inférence de réseau d'interactions pour les données de comptage. Le modèle PLN appartient à la famille des modèles à variables latentes, car il modélise l'abondance des entités en utilisant une distribution de Poisson paramétrée par une variable aléatoire gaussienne latente.

Soit $\mathbf{X} \in \mathbb{R}^{n \times d}$ une matrice d'abondance de $d$ entités parmi $n$ sites. Le modèle PLN repose sur les hypothèses suivantes.

**H1**
- $\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ où $\boldsymbol{\mu}_i \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$.
- $(\mathbf{Z}_i, \mathbf{X}_i)_{1 \leq i \leq n}$ sont indépendants, et conditionnellement à $\mathbf{Z}$ et $X_k$, $1 \leq k \neq j \leq d$, $X_j$ dépend de $Z_j$ uniquement et $X_j \sim \mathcal{P}\left(\mathrm{e}^{Z_j}\right)$.

On note $p_{\mathbf{Z},\boldsymbol{\theta}}$ la densité de $\mathbf{Z}$ où $\theta = \{\boldsymbol{\Sigma}, (\boldsymbol{\mu}_i)_{1 \leq i \leq n}\}$. Les modèles partiellement observés comme PLN font généralement usage de l'algorithme EM pour inférer le paramètre $\boldsymbol{\theta}$. Cependant, l'étape E nécessite le calcul d'une espérance conditionnelle sous la distribution $p_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X})$ qui n'est pas explicite dans le cadre de PLN. Les auteurs de [Chiquet et al., 2021] suggèrent une approximation variationnelle de l'étape E en optimisant une quantité sous-optimale pour la log-vraisemblance appelée ELBO [Kingma et al., 2019]:

$$\mathrm{ELBO}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbb{E}_{q_{\boldsymbol{\varphi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}[q_{\boldsymbol{\varphi}}(\cdot|\mathbf{X}), p_{\mathbf{Z},\boldsymbol{\theta}}(\cdot)] \; . \tag{1}$$

L'approximation variationnelle choisie par [Chiquet et al., 2021] s'associe à l'approche champ moyen :

$$q_{\boldsymbol{\varphi}}(\mathbf{Z}_i|\mathbf{X}_i) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{Z}_i; \mathbf{m}_i, \mathbf{S}_i) \; ,$$

où $\mathcal{N}(\cdot; \mathbf{m}_i, \mathbf{S}_i)$ est la densité de la loi gaussienne de moyenne $\mathbf{m}_i \in \mathbb{R}^d$ et de variance $\mathbf{S}_i = \mathrm{diag}(\mathbf{s}_i)$ avec $\mathbf{s}_i \in \mathbb{R}^d$ et où $\boldsymbol{\varphi}$ représente tous les paramètres variationnels. En conséquence, l'EM variationnel génère une séquence de paramètres $(\boldsymbol{\theta}^h, \varphi^h)_{1 \leq h \leq N}$ obtenus en réalisant une optimisation alternée consistant d'abord à optimiser l'ELBO relativement à $\boldsymbol{\varphi}$ tel que

$$\begin{aligned} \boldsymbol{\varphi}^{h+1} = \quad & \arg\max_{\boldsymbol{\varphi}} \mathrm{ELBO}(\boldsymbol{\theta}^h, \boldsymbol{\varphi}) \\ & \text{t.q.} \quad \forall i \in \{1, \ldots, n\}, \mathbf{S}_i \succ 0 \; . \end{aligned}$$

L'approximation variationnelle choisie rend l'ELBO explicite, permettant ainsi une optimisation directe de $\boldsymbol{\varphi}$ [Chiquet et al., 2021, Chiquet et al., 2019]. Il devient ensuite possible

d'effectuer l'étape E en remplaçant la vraie loi *a posteriori* $p_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X})$ par $q_{\boldsymbol{\varphi}^{h+1}}(\mathbf{Z}|\mathbf{X})$, ce qui permet de calculer l'étape M pour obtenir $\boldsymbol{\theta}^{h+1}$, et ainsi de suite jusqu'à convergence.

Ce modèle s'est avéré très efficace pour adapter aux données de comptage un large spectre d'outils d'analyse multivariée tels que l'ACP, l'analyse discriminante et l'inférence de réseau. L'application d'intérêt pour notre problème est l'inférence de réseau, qui nous permettrait de découvrir les interactions fonctionnelles entre les bactéries du microbiome intestinal. Le problème d'inférence de réseau s'accompagne également d'une architecture informée par la parcimonie suggérée par [Chiquet et al., 2019], en introduisant une pénalisation de type LASSO sur la matrice de précision $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

Enfin, les auteurs soulignent le rôle clé des covariables et des offsets dans le problème d'inférence de graphe qui peuvent être intégrés dans la moyenne du modèle latent.

## 2.2 Top-Down PLNTree

Considérons maintenant que les données de comptage proviennent d'un arbre taxonomique connu, noté $\mathcal{T}$. Un exemple d'échantillon est fourni ci-dessous.

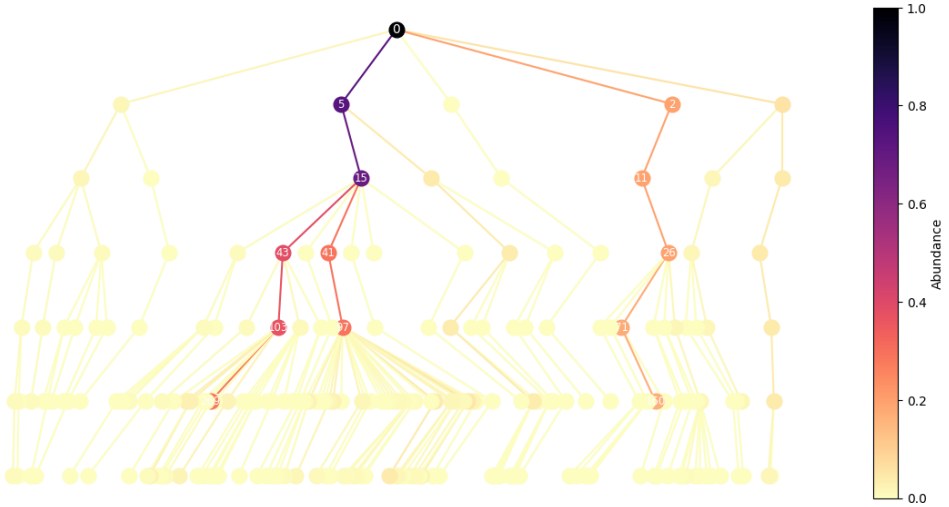



Figure 1: Exemple d'échantillon d'abondance de taxons (coloré selon l'abondance relative)

En supposant que $\mathcal{T}$ comporte $L$ couches, $K_\ell$ nœuds à la couche $\ell$, nous désignons les échantillons d'abondance de taxons par

$$\mathbf{X} = (\mathbf{X}^1, \ldots, \mathbf{X}^L) \,,$$

où à la couche $\ell$, $\mathbf{X}^\ell = (\mathbf{X}_1^\ell, \ldots, \mathbf{X}_{K_\ell}^\ell)$. Pour un nœud donné $X_k^\ell$, nous introduisons la notation $\mathcal{C}\left(X_k^\ell\right)$ qui est l'ensemble des enfants de ce nœud. La notation $\mathcal{C}_\ell^k$ renvoie à l'ensemble des indices des enfants du nœud $X_k^\ell$.

Pour modéliser la structure séquentielle de l'arbre à travers ses couches, nous introduisons une dynamique de Markov dans l'espace latent et dans l'espace observé tout en restant dans le cadre du modèle PLN. Supposons que l'ensemble d'hypothèses suivant soit vérifié :

**H2**
- Les $(\mathbf{Z}_i, \mathbf{X}_i)_{1 \leq i \leq n}$ sont indépendants, et conditionnellement à $\mathbf{Z}$ et $\mathbf{X} \backslash \mathcal{C}\left(X_k^\ell\right), 1 \leq \ell \leq L, 1 \leq k \leq K_\ell$, l'ensemble des variables $\mathcal{C}\left(X_k^\ell\right)$ ne dépend que de $\mathcal{C}\left(Z_k^\ell\right)$ et de $X_k^\ell$.
- Dynamique Markovienne latente (de haut en bas) :

$$\begin{aligned} \mathbf{Z}^1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \ , \\ \mathbf{Z}^{\ell+1} &\sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}_{\ell+1}}(\mathbf{Z}^\ell), \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\ell+1}}(\mathbf{Z}^\ell)) \ , \end{aligned}$$

  où $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\ell+1}}(.)$ et $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\ell+1}}(.)$ désignent des fonctions arbitraires paramétrées par $\boldsymbol{\theta}_{\ell+1}$.
- Dynamique observée contrainte :

$$\begin{aligned} \mathbf{X}^1 &\sim \mathcal{P}\left(\mathrm{e}^{\mathbf{Z}^1}\right) \ , \\ \mathcal{C}\left(X_k^\ell\right) &\sim \mathcal{P}\left(\mathrm{e}^{\mathcal{C}\left(Z_k^\ell\right)} \middle| \sum_{j \in \mathcal{C}_k^\ell} X_j^{\ell+1} = X_k^\ell\right) \ . \end{aligned}$$

**Approximation Markov Backward.** Comme dans le cas du modèle PLN, nous n'avons pas accès à un estimateur du maximum de vraisemblance. Au lieu de cela, comme suggéré par [Chiquet et al., 2021] dans le modèle PLN, nous effectuons une approximation variationnelle pour estimer la distribution a posteriori non-explicite. Sous H2, la distribution de lissage visée $p_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X})$ est une chaîne de Markov "backward" :

$$p_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X}) = p_{\boldsymbol{\theta}}(\mathbf{Z}^L|\mathbf{X}^{1:L}) \prod_{\ell=1}^{L-1} p_{\boldsymbol{\theta}}(\mathbf{Z}^\ell|\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}) \ .$$

Puisque nous approchons la quantité ci-dessus à l'aide d'une approximation variationnelle, nous suggérons une famille variationnelle qui tient compte de la structure "backward" de la vraie distribution de lissage :

**H3** La loi variationnelle "backward" s'écrit :

$$q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}) = q_{\boldsymbol{\varphi}^L}(\mathbf{Z}^L|\mathbf{X}^{1:L}) \prod_{\ell=1}^{L-1} q_{\boldsymbol{\varphi}^\ell}(\mathbf{Z}^\ell|\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}) \ ,$$

où

- $q_{\boldsymbol{\varphi}^L}(\mathbf{Z}^L|\mathbf{X}^{1:L}) = \mathcal{N}(\mathbf{Z}^L; \mathbf{m}_{\boldsymbol{\varphi}^L}(\mathbf{X}^{1:L}), \mathbf{S}_{\boldsymbol{\varphi}^L}(\mathbf{X}^{1:L}))$
- $q_{\boldsymbol{\varphi}^\ell}(\mathbf{Z}^\ell|\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}) = \mathcal{N}(\mathbf{Z}^\ell; \mathbf{m}_{\boldsymbol{\varphi}^\ell}(\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}), \mathbf{S}_{\boldsymbol{\varphi}^\ell}(\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}))$

– Pour tout $\ell$, $\mathbf{S}_{\boldsymbol{\varphi}^\ell}$ produit une matrice diagonale qui peut être représentée par sa diagonale.

Manipuler les entrées des réseaux employés en pratique est une tâche difficile, sans formulation évidente. Nous suggérons une configuration expérimentale amortie utilisant un réseau de neurones récurrent, inspirée de récents travaux [Chagneux et al., 2022], de sorte que $\mathbf{X}^{1:\ell}$ soit incorporé sous une forme contractée pour neutraliser la nature expensive de la chaîne.

# 3 Résultats numériques

Dans cette section, nous visons à montrer les performances empiriques du modèle taxa-PLN, que nous souhaitons comparer à celles de PLN afin de mesurer l'intérêt d'intégrer la taxonomie au modèle. Considérons une étude pratique d'une cohorte de 732 patients atteints de la maladie de Crohn.

**Approximation champ moyen.** L'approximation variationnelle régulièrement utilisée dans la littérature est l'approximation champ moyen (mean-field) [Kingma et al., 2019], qui s'exprime comme suit.

**H4** L'approximation fonctionnelle du champ moyen Gaussien est donnée par

$$q_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{X}) = \prod_{\ell=1}^{L} q_{\boldsymbol{\varphi}^\ell}(\mathbf{Z}^\ell|\mathbf{X}^\ell) \ ,$$

où $q_{\boldsymbol{\varphi}^\ell}\left(\mathbf{Z}^\ell|\mathbf{X}^\ell\right) \sim \mathcal{N}(\mathbf{m}_{\boldsymbol{\psi}_\ell}(\mathbf{X}^\ell), \mathbf{S}_{\boldsymbol{\psi}^\ell}(\mathbf{X}^\ell))$, tel que $\mathbf{m}_{\boldsymbol{\psi}_\ell}(\mathbf{X}^\ell) \in \mathbb{R}^{K_\ell}$, et $\mathbf{S}_{\boldsymbol{\psi}_\ell}(\mathbf{X}^\ell) = \mathrm{diag}(\mathbf{s}_{\boldsymbol{\psi}_\ell}(\mathbf{X}^\ell))$ avec $\mathbf{s}_{\boldsymbol{\psi}_\ell}(\mathbf{X}^\ell) \in \mathbb{R}_+^{K_\ell}$ .

En utilisant l'approximation champ moyen, nous obtenons un premier candidat de comparaison avec l'approximation "backward" de PLNTree. Bien qu'il soit plus facile d'entraîner un modèle champ moyen, cette approximation peut cependant restreindre trop fortement le choix de la famille variationnelle par rapport à la vraie distribution *a posteriori*.

**Benchmarks génératifs.** Nous commençons par évaluer la capacité du modèle à générer des échantillons de qualité, dans le sens où ils sont proches de l'échantillon initial qui a servi à apprendre les paramètres du modèle. Pour ce faire, nous apprenons le modèle PLNtree sur l'ensemble des données et nous générons des données artificielles de taxa-abondance (MF : Champ moyen, Amortized : Chaîne de Markov backward amortie). Nous procédons à une procédure similaire en utilisant le modèle PLN à chaque couche. Toutefois, empiler les comptages obtenus via PLN ne fournit pas des échantillons plausibles, car ils ne respectent pas les contraintes de taxa-abondance. Par conséquent, nous fournissons également une variante PLN (fill) qui permet de générer des échantillons de taxa-abondance valides en

partant uniquement de la dernière couche, puis en remontant l'arbre par les contraintes déterministes de la taxonomie.

Les critères de comparaison des générations correspondent aux mesures de diversité alpha couramment utilisées dans l'analyse du microbiote, à savoir, l'entropie de Shannon, Inverse Simpson ou encore l'indice Chao1 [Chao, 1984].

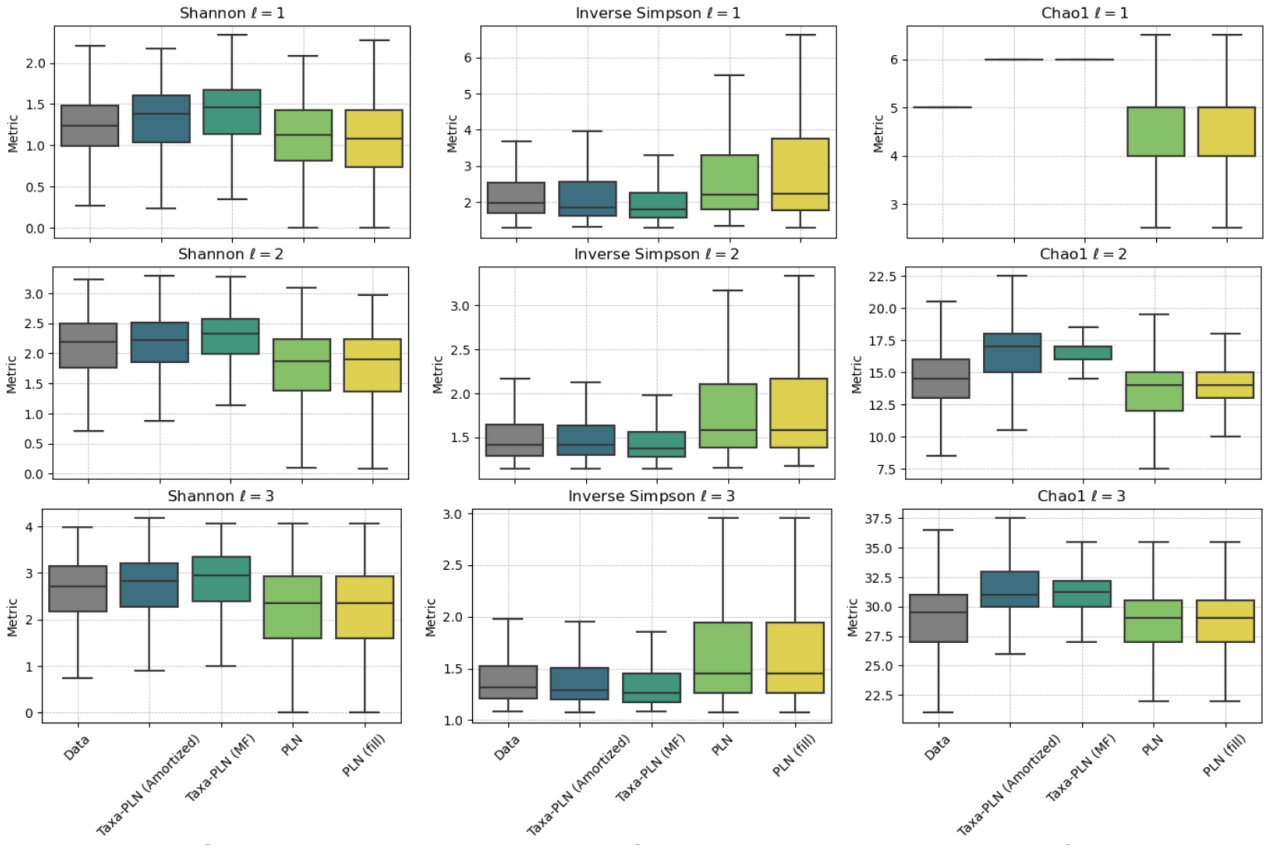



Figure 2: Benchmarks de génération de diversité alpha sur Suivitheque pour chaque couche.

En regardant la figure 2, il semble que l'exploitation de l'information taxonomique ait amélioré la qualité générative du modèle. En regardant l'indice Chao1, on peut supposer que les échantillons générés souffrent de comptes spurieux (comptes proches de 0) qui augmentent la diversité Chao1, ouvrant la voie à d'autres améliorations. Les figures 2 et 3 ainsi que le tableau 1 montrent également les améliorations apportées par l'inférence backward amortie par rapport à l'approche champ moyen.

| | Taxa-PLN (MF) | Taxa-PLN (Amortized) | PLN | PLN (fill) |
|---|---|---|---|---|
| $\bar{\rho}$ | $0.96 \pm 0.13$ | $0.99 \pm 0.08$ | $0.99 \pm 0.04$ | $0.99 \pm 0.04$ |

Table 1: Corrélation moyenne entre les échantillons originaux et après encodage puis décodage.

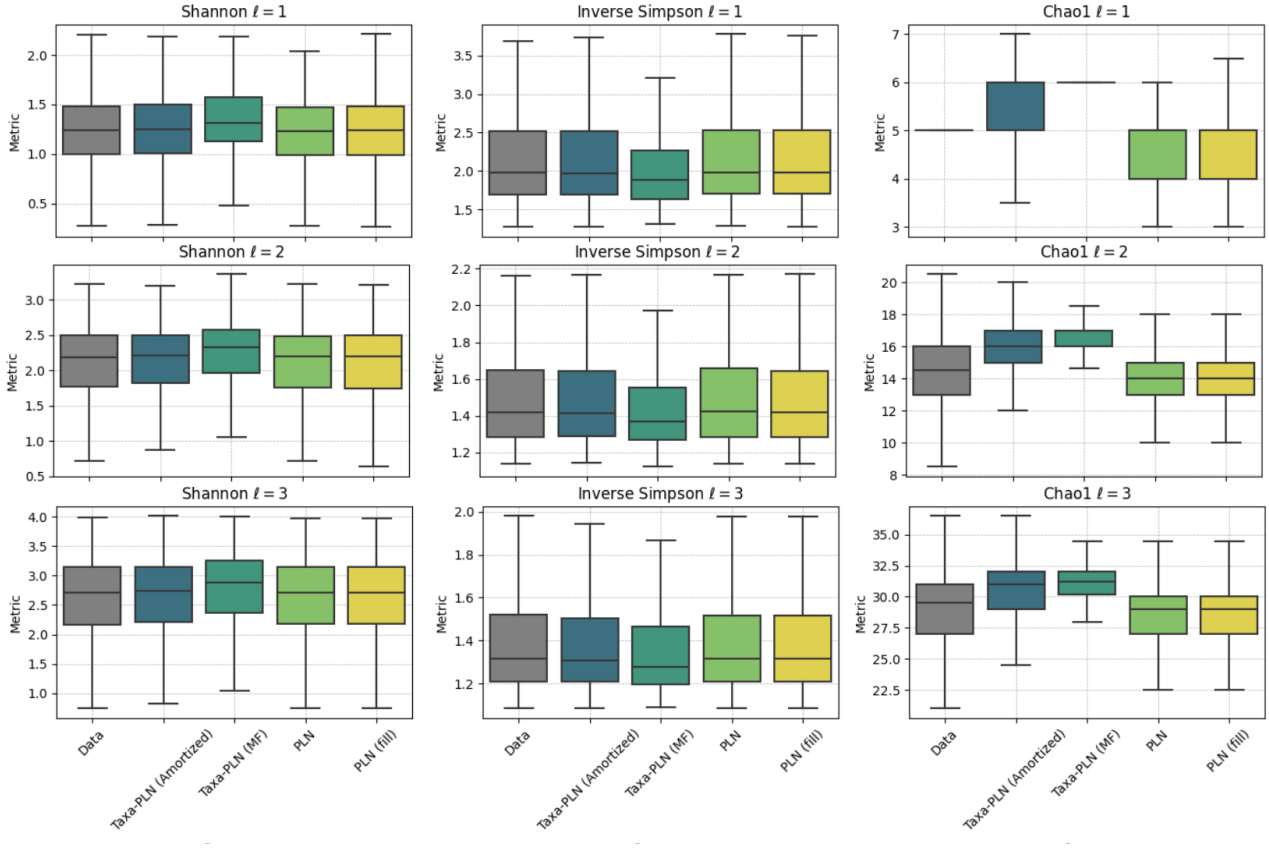



Figure 3: Benchmarks d'encodage de la diversité alpha sur Suivitheque pour chaque couche.

# 4 Conclusion

Le modèle PLNtree apparaît comme un modèle pertinent pour inférer les interactions microbiennes au sein du microbiome intestinal. Les gains de performance en alpha diversité proposés par PLNtree sur la tâche de génération montrent l'intérêt d'inclure l'information taxonomique dans l'inférence des interaction microbiennes. De plus, l'introduction d'une chaîne de Markov "backward" pour l'approximation variationnelle semble améliorer la précision du modèle, en montrant des résultats supérieurs au cas champ moyen. La perspective d'obtenir des garanties statistiques grâce à cette approximation ajoute une dimension prometteuse aux contributions potentielles du modèle PLNtree dans la compréhension des écosystèmes microbiens complexes.

# Bibliography

# References

[Aitchison and Ho, 1989] Aitchison, J. and Ho, C. (1989). The multivariate poisson-log normal distribution. *Biometrika*, 76(4):643–653.

[Altenbuchinger et al., 2020] Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H. J., and Zacharias, H. U. (2020). Gaussian and mixed graphical models as (multi-) omics data analysis tools. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194418.

[Chagneux et al., 2022] Chagneux, M., Gassiat, É., Gloaguen, P., and Corff, S. L. (2022). Amortized backward variational inference in nonlinear state-space models. *arXiv preprint arXiv:2206.00319*.

[Chao, 1984] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270.

[Chiquet et al., 2021] Chiquet, J., Mariadassou, M., and Robin, S. (2021). The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:588292.

[Chiquet et al., 2019] Chiquet, J., Robin, S., and Mariadassou, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171. PMLR.

[Inouye et al., 2017] Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398.

[Julien et al., 2022] Julien, C., Anakok, E., Treton, X., Nachury, M., Nancey, S., Buisson, A., Fumery, M., Filippi, J., Maggiori, L., Panis, Y., et al. (2022). Impact of the ileal microbiota on surgical site infections in crohn's disease: a nationwide prospective cohort. *Journal of Crohn's and Colitis*, 16(8):1211–1221.

[Kingma et al., 2019] Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

# Pólya urn and multivariate birth-death processes under neutral theory of biodiversity

Jean Peyhardi [1] & Fabien Laroche [2], Frédéric Mortier [3]

[1] *IMAG, University of Montpellier, CNRS, Montpellier, France*
[2] *MR DYNAFOR, INP de Toulouse, INRAE, Auzeville Tolosane, France*
[3] *CIRAD, UPR Forêts et Sociétés, F-34398 Montpellier, France*

**Résumé.** La famille des modèles joints de distribution d'espèces est devenue un outil statistique utile pour analyser les données multivariées d'abondance d'espèces. La théorie neutre unifiée de la biodiversité introduite par Hubbell et al. (2001) souligne l'importance des processus stochastiques dans la structure des communautés écologiques. Sous l'hypothèse de taille de communauté fixée et l'hypothèse de neutralité (les individus de différentes espèces sont écologiquement identiques), il a montré que la distribution stationnaire d'un processus multivarié de naissance/mort est une Dirichlet multinomiale. Etienne et al. (2007) ont relaché l'hypothèse de somme fixe et ont également trouvé une distribution Dirichlet multinomiale. Mais ils ont ajouté l'hypothèse d'indépendance entre espèces. Nous proposons de généraliser cette approche, en s'affranchissant de l'hypothèse d'indépendance et en considérant la famille générale des distributions de Pólya. L'indépendance est donc une conséquence des hypothèses paramétriques faites sur les taux de natalité/mortalité et non plus une hypothèse. Nous mettons également en évidence neuf distributions remarquables de Pólya (somme aléatoire) qui sont stables par marginalisation. Nous donnons la forme paramétrique des taux de saut conduisant à ces neuf distributions, en incluant le cas particulier de la Dirichlet multinomiale (somme binomiale négative), obtenu par Etienne et al. (2007).

**Mots-clés.** Theorie neutre, Modèle joint de distribution d'espèces, Urnes de Pólya

**Abstract.** The family of joint species distribution models (JSDMs) has emerged as a useful statistical tool to analyse multivariate species abundance data. The unified neutral theory of biodiversity introduced by Hubbell et al. (2001) emphasizes the importance of stochastic processes in ecological community structure. Under the zero-sum assumption (fixed community size) and the neutrality assumption (individuals of different species are ecologically identical) he showed that the stationary distribution of a multivariate birth and death process is a Dirichlet multinomial. Etienne et al. (2007) relaxed the zero-sum assumption and also found a Dirichlet multinomial distribution. But they added the assumption of independence between species. We propose to generalize this approach in two ways, by relaxing the independence assumption and by considering the enlarged family of Pólya distributions. Independence is thus a consequence of parametric assumption made on birth/death rates and not a necessary assumption. We also highlight nine remarkable Pólya splitting distributions that are closed under all marginalizations. Then we give the parameteric form of jumping rates leading to these nine distributions, recovering the special case Dirichlet multinomial split with negative binomial sum obtained by Etienne et al. (2007).

**Keywords.** Neutral theory, Joint Species Distribution Model, Pólya urn

# 1 Biological context

Biodiversity is not only determined by the number of different entities (species in community ecology or alleles in population genetics), but also by the abundance of these entities (Magurran, 2021). Many ecological questions require the joint analysis of abundances collected simultaneously across many taxonomic groups. The family of joint species distribution models (JSDMs) has then emerged as a useful statistical tool to analyse such data. When species abundance are collected as count data (instead of presence/absence) a JSDMs turns out to be a multivariate count distribution of a random vector $\boldsymbol{N} = (N_1, \ldots, N_J)$ where $J$ denotes the number of species. The unified neutral theory of biodiversity introduced by Hubbell et al. (2001) emphasizes the importance of stochastic processes in ecological community structure, and has challenged the traditional niche-based view of ecology. It has challenged classical theories of species diversity by showing that patterns of species diversity similar to those observed in nature can be obtained from an extremely simplified model of community dynamics where each dying individual is immediately replaced by a new individual (zero-sum) and all individuals of all species are ecologically identical (neutrality). From a mathematical point of view, focus is made on the stationary distribution of a multivariate birth and death process $\boldsymbol{N}(t) = (N_1(t), \ldots, N_J(t))$, where $N_j(t)$ is the abundance of species $j$ at time $t$. Under the zero-sum assumption (i.e., fixed sum $|\boldsymbol{N}(t)| = n$) - and mild hypothesis on birth and death rates - Hubbell et al. (2001) showed that the multivariate distribution of $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$ (at equilibrium) is a Dirichlet multinomial distribution $\mathcal{DM}_n(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$. Etienne et al. (2007) proposed to relax the zero-sum assumption in order to obtain a more realistic model and they also find a Dirichlet multinomial distribution at equilibrium for the conditional distribution of $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$. But they added the assumption of independence between species to obtain this result since in this case it is sufficient to study the abundance distribution for each species and then calculate the product. In the present work we propose to generalize this approach in two ways, by relaxing the independence assumption and by considering the enlarged family of multivariate Pólya distributions. This work is based on both papers of Peyhardi (2023) and Peyhardi et al. (2024).

# 2 Pólya splitting distributions

The class of Pólya splitting distributions has been introduced by Peyhardi et al. (2021) as compound distributions $\boldsymbol{N} \sim \mathcal{P}_n^{[c]}(\boldsymbol{\theta}) \wedge_n \mathcal{L}$, meaning that $|\boldsymbol{N}|$ follows the univariate distribution $\mathcal{L}$ and $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$ follows the multivariate Pólya distribution. Let us briefly recall the definition of a multivariate Pólya distribution in terms of urn models.

## 2.1 Pólya urn model

One urn initially contains $\theta_j$ balls of the color $j$ for $j = 1, \ldots, J$. At each draw, one ball is drawn at random and then replaced with $c$ additional balls of the same color, where $c \in \{-1, 0, 1\}$. This procedure is repeated $n$ times and focus is made on the multivariate

count $\boldsymbol{N} = (N_1, \ldots, N_J)$ of drawn balls for each color. Knowing the number $n$ of draws, the conditional count distribution of $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$ is known as the multivariate Pólya distribution, denoted by $\mathcal{P}_n^{[c]}(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta_c^J$ (where $\Theta = \mathbb{N}$ for $c = -1$ and $\Theta = \mathbb{R}_+$ otherwise). As expressed by Peyhardi (2023), if we denote

$$a_\theta^{[c]}(n) = \frac{\prod_{k=0}^{n-1}(\theta + ck)}{n!} \mathbb{1}_{\theta + cn \geq 0},$$

then the probability mass function (pmf) of a multivariate Pólya distribution $\mathcal{P}_n^{[c]}(\boldsymbol{\theta})$ takes the following form

$$P_{|\boldsymbol{N}|=n}(\boldsymbol{N} = \boldsymbol{n}) = \frac{1}{a_{|\boldsymbol{\theta}|}^{[c]}(n)} \prod_{j=1}^{J} a_{\theta_j}^{[c]}(n_j).$$

The mulyivariate Pólya distribution turns out to be the:

- multivariate hypergeometric distribution when $c = -1$ (i.e., without replacement), denoted by $\mathcal{H}_{\Delta_n}(\boldsymbol{\theta})$ with pmf

$$P_{|\boldsymbol{N}|=n}(\boldsymbol{N} = \boldsymbol{n}) = \frac{1}{\binom{|\boldsymbol{\theta}|}{n}} \prod_{j=1}^{J} \binom{\theta_j}{n_j}.$$

- multinomial distribution when $c = 0$ (i.e., with replacement meaning independent draws), denoted by $\mathcal{M}_{\Delta_n}(\boldsymbol{\pi})$ with pmf

$$P_{|\boldsymbol{N}|=n}(\boldsymbol{N} = \boldsymbol{n}) = \frac{1}{|\boldsymbol{\theta}|/n!} \prod_{j=1}^{J} \frac{\theta_j^{n_j}}{n_j!}.$$

- Dirichlet-multinomial distribution when $c = 1$ (i.e., with reinforcement), denoted by $\mathcal{DM}_{\Delta_n}(\boldsymbol{\theta})$ with pmf

$$P_{|\boldsymbol{N}|=n}(\boldsymbol{N} = \boldsymbol{n}) = \frac{1}{\binom{n+|\boldsymbol{\theta}|-1}{n}} \prod_{j=1}^{J} \binom{n_j + \theta_j - 1}{n_j}.$$

## 2.2 Remarkable Pólya splitting distributions

Properties of Pólya splitting distributions are related to the choice of the sum distribution $\mathcal{L}$. For instance the covariance between $N_i$ and $N_j$ (with $(i, j) \in \{1, \ldots, J\}^2$ and $i \neq j$) is given by

$$\mathrm{Cov}(N_i, N_j) = \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2(|\boldsymbol{\theta}| + c)} \left[(\mu_2 - \mu_1^2)|\boldsymbol{\theta}| - c\mu_{(1)}^2\right], \tag{1}$$

where $\mu_k$ is the factorial moment of order $k$ of the sum distribution $\mathcal{L}$. It implies that the sign of covariance between any pair $(i, j)$ is driven by the moments of the sum; see Table 1 for some examples. In fact it is possible to characterize the probabilistic graphical model (PGM) of a Pólya splting distribution according $\mathcal{L}$; see Theorem 1. We first need to introduce three remarkable choices for the sum distribution $\mathcal{L}$ (see (Peyhardi, 2023) for details):

1. The (univariate) Pólya distribution $\mathcal{P}_n^{[c]}(\theta, \gamma)$ with pmf

$$P(X = k) = \frac{a_\theta^{[c]}(k) a_\gamma^{[c]}(n-k)}{a_{\theta+\gamma}^{[c]}(n)} \mathbb{1}_{k \leq n},$$

   It includes the hypergeometric ($c = -1$), binomial ($c = 0$) and beta binomial ($c = 1$).

2. The power series distribution $\mathcal{PS}^{[c]}(\theta, \alpha)$ with pmf

$$P(X = k) = \frac{a_\theta^{[c]}(k) \alpha^k}{h_\theta(\alpha)}$$

   It includes the binomial ($c = -1$), Poisson ($c = 0$) and negative binomial ($c = 1$).

3. The inverse Pólya distribution $\mathcal{IP}^{[c]}(r; \theta, \gamma)$ with pmf

$$P(X = k) = \frac{r}{k+r} \frac{a_\theta^{[c]}(k) a_\gamma^{[c]}(r)}{a_{\theta+\gamma}^{[c]}(k+r)}$$

   It includes the beta binomial ($c = -1$), negative binomial ($c = 0$) and negative beta binomial ($c = 1$).

**Theorem 1 ((Peyhardi, 2023))** *The PGM of a Pólya splitting distribution* $\mathcal{P}_n^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}$ *is:*

- *empty if and only if* $\mathcal{L} = \mathcal{PS}^{[c]}(|\boldsymbol{\theta}|, \alpha)$
- *complete otherwise*

This theorem generalizes the result of Bol'shev (1965) and those of Rao and Janardan (1984) that concerns the bivariate case ($J = 2$) only for some thinning operators.

**Theorem 2 ((Peyhardi, 2023))** *The Pólya, the power series and the inverse Pólya distributions are closed under the Pólya thinning operator* $\mathcal{P}_n^{[c]}(\theta, \gamma) \wedge_n (\cdot)$ *when their parameter respects the additive constraint. More precisely we have the following distributions equalities:*

1. $\mathcal{P}_n^{[c]}(\theta, \gamma) \underset{n}{\wedge} \mathcal{P}_m^{[c]}(\theta + \gamma, \lambda) = \mathcal{P}_m^{[c]}(\theta, \gamma + \lambda)$,
2. $\mathcal{P}_n^{[c]}(\theta, \gamma) \underset{n}{\wedge} \mathcal{PS}^{[c]}(\theta + \gamma, \alpha) = \mathcal{PS}^{[c]}(\theta, \alpha)$,
3. $\mathcal{P}_n^{[c]}(\theta, \gamma) \underset{n}{\wedge} \mathcal{IP}^{[c]}(r; \theta + \gamma, \lambda) = \mathcal{IP}^{[c]}(r; \theta, \lambda)$,

*for all* $(\theta, \gamma, \lambda) \in \Theta^3$, $\alpha \in (0, R)$ *and* $r \in (0, \infty)$.

As corollary of Theorem 2, based on specific case $c = 0$, the result of Rao (1965) is recovered.

**Corollary 1 ((Rao, 1965))** *The binomial, Poisson and negative binomial distribution are closed under the binomial thinning operation.*

1. *binomial:* $\mathcal{B}_n(\pi) \underset{n}{\wedge} \mathcal{B}_m(p) = \mathcal{B}_m(p')$ *where* $p' = \pi p \in (0,1)$,
2. *Poisson:* $\mathcal{B}_n(\pi) \underset{n}{\wedge} \mathcal{P}(\lambda) = \mathcal{P}(\lambda')$ *where* $\lambda' = \pi\lambda \in (0,\infty)$,
3. *negative binomial:* $\mathcal{B}_n(\pi) \underset{n}{\wedge} \mathcal{NB}(r,p) = \mathcal{NB}(r,p')$ *where* $p' := \frac{\pi p}{\pi p + 1 - p} \in (0,1)$.

Theorem 2 allows us to build nine remarkable Pólya splitting distributions (see Table 1) that are closed under all marginalizations. According to equation (1) the sign of covariance is easily obtained in this nine examples.

| Sum \ Split | Hypergeometric $c=-1$ | Multinomial $c=0$ | Dirichlet multinomial $c=1$ | Covariance sign |
|---|---|---|---|---|
| Pólya | $\mathcal{H}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{H}_m(\lvert\boldsymbol{\theta}\rvert,\gamma)$ | $\mathcal{M}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{B}_m(p)$ | $\mathcal{DM}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{B}_m(\lvert\boldsymbol{\theta}\rvert,\gamma)$ | negative |
| Power series | $\mathcal{H}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{B}_{\lvert\boldsymbol{\theta}\rvert}(p)$ | $\mathcal{M}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{P}(\lambda)$ | $\mathcal{DM}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{NB}(\lvert\boldsymbol{\theta}\rvert,p)$ | null |
| Inverse Pólya | $\mathcal{H}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{B}_{\lvert\boldsymbol{\theta}\rvert}(a,b)$ | $\mathcal{M}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{NB}(a,p)$ | $\mathcal{DM}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{NB}(\lvert\boldsymbol{\theta}\rvert,a,b)$ | positive |

Table 1: Nine remarkable Pólya splitting distributions with different split distributions (columns) and different sum distributions (rows).

# 3 Multivariate birth-death processes under neutrality

We will now exhibit the birth and death rates assumptions that lead to the multivariate Pólya distribution at equibrium. The master equation describing the behaviour of the multivariate jump process $\boldsymbol{N}(t)$ is given by

$$\frac{\partial p_{\boldsymbol{n}}(t)}{\partial t} = \sum_{j=1}^{J} p_{\boldsymbol{n}-\boldsymbol{e}_j}(t) q_j^-(\boldsymbol{n}-\boldsymbol{e}_j) + p_{\boldsymbol{n}+\boldsymbol{e}_j}(t) q_j^+(\boldsymbol{n}+\boldsymbol{e}_j) - p_{\boldsymbol{n}+\boldsymbol{e}_j}(t)\{q_j^-(\boldsymbol{n}) + q_j^+(\boldsymbol{n})\}$$

where $q_j^-(\boldsymbol{n})$ (resp. $q_j^+(\boldsymbol{n})$) denotes the jumping rate from $\boldsymbol{n}$ to $\boldsymbol{n}-\boldsymbol{e}_j$ (resp. to $\boldsymbol{n}+\boldsymbol{e}_j$), $\boldsymbol{e_j}$ denotes the indicator vector of the $j_{\text{th}}$ element and $p_{\boldsymbol{n}}(t) = P\{\boldsymbol{N}(t) = \boldsymbol{n}\}$ (resp. $p_{\boldsymbol{n}} = P(\boldsymbol{N} = \boldsymbol{n})$) denotes the pmf at time $t$ (resp. the pmf at stationary state). Assume that there exists some parameters $c \in \{-1,0,1\}$ and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_J) \in \Theta_c^J$ and two non-negative functions $s^+$ and $s^-$ such that the birth and death rates have the following form

$$\begin{aligned} q_j^+(\boldsymbol{n}) &= s^+(\lvert\boldsymbol{n}\rvert)(\theta_j + c n_j)\mathbb{1}_{\theta_j + c n_j \geq 0}, \\ q_j^-(\boldsymbol{n}) &= s^-(\lvert\boldsymbol{n}\rvert) n_j. \end{aligned} \tag{2}$$

The birth-deah rate $q_j(\boldsymbol{n}) := q_j^+(\boldsymbol{n})/q_j^-(\boldsymbol{n}+\boldsymbol{e_j})$ thus becomes

$$q_j(\boldsymbol{n}) = s(|\boldsymbol{n}|)\frac{\theta_j + cn_j}{n_j+1}\mathbb{1}_{\theta_j+cn_j\geq 0} \tag{3}$$

where $s(n) = \frac{s^+(n)}{s^-(n+1)}$ for all $n \in \mathbb{N}$. It can be seen that this parametric assumption (3) respects the Kolmogorov's criterion $q_i(\boldsymbol{n})q_j(\boldsymbol{n}+\boldsymbol{e_i}) = q_j(\boldsymbol{n})q_i(\boldsymbol{n}+\boldsymbol{e_j})$, and thus leads to a reversible process. Remarking that $\prod_{k=0}^{n-1}\frac{\theta+ck}{k+1} = a_\theta^{[c]}(n)$ we add the following assumption on $s(n)$ in order to obtain a well defined stationary distribution:

$$\sum_{n\geq 0} a_\theta^{[c]}(n)\prod_{k=0}^{n-1} s(k) < \infty. \tag{4}$$

**Theorem 3 ((Peyhardi et al., 2024))** *Assume that the hypothesis* (3) *and* (4) *hold then*

- *the stationary distribution of* $\boldsymbol{N}(t)$ *is the Pólya splitting distribution* $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}$
- $\mathcal{L}$ *is the stationary distribution of a univariate process with birth/death ratio equal to* $q(n) = s(n)r_{|\boldsymbol{\theta}|}^{[c]}(n)$*, more precisely we have* $P(|\boldsymbol{N}| = n) \propto a_{|\boldsymbol{\theta}|}^{[c]}(n)\prod_{k=0}^{n-1} s(k)$.

Theorem 1 characterizes the independence between variable $N_1, \ldots, N_J$ through the distribution of the sum $|\boldsymbol{N}|$. In terms of multivariate birth-death process the sum distribution (at equilibrium) is governed by the quantity $s(|\boldsymbol{n}|)$ that appears in jumping rates form (4). The assumption $s(|\boldsymbol{n}|) = \alpha$ thus leads to the power series distribution $|\boldsymbol{N}| \sim \mathcal{PS}^{[c]}(|\boldsymbol{\theta}|, \alpha)$ and then to independence between species. Independence is a consequence of parametric assumption made on birth and death rates and not a necessary assumption *per se*, contrary to what was posited by other authors (Etienne et al., 2007).

Theorem 2 allows us to build nine remarkable Pólya splitting distributions (see Table 1) that are closed under all marginalizations. Using our framework, the parameteric form of jumping rates leading to these distributions are easily obtained (see Table 2). Moreover it has been shown that covariance between two species abundancies $\mathrm{Cov}(N_i, N_j)$ have the same sign for any pair of species and are negative (resp. null or positive) when the sum distribution is the Pólya (resp. the power series or the inverse Pólya).

Using Theorem 3, the parameteric form of jumping rates leading to these distributions are easily obtained (see Table 2). Overall, we advocate that Pólya-splitting distribution should become a part of the classic toolbox for the analysis of multivariate count data in ecology, providing alternative approaches to JSDM framework. Moreover, explanatory variables are easily taken into account in this framework (with a decomposable likelihood according to the compound distribution) leading to competitve statistical models compare to classical Poisson log-normal model that have not undelying process interpretation.

| Sum \ Split | Hypergeometric $c=-1$ | Multinomial $c=0$ | Dirichlet multinomial $c=1$ | Covariance sign |
|---|---|---|---|---|
| Pólya | $\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma-m+\lvert\boldsymbol{n}\rvert+1}$ | $\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma}$ | $\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma+m-\lvert\boldsymbol{n}\rvert-1}$ | negative |
| Power series | $\alpha$ | $\alpha$ | $\alpha$ | null |
| Inverse Pólya | $\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b-\lvert\boldsymbol{n}\rvert-1}$ | $\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b}$ | $\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b+a+\lvert\boldsymbol{n}\rvert+1}$ | positive |

Table 2: Parametric form of $s(|\boldsymbol{n}|)$ and thus of jumping rates $q_j(\boldsymbol{n}) = q_j^+(\boldsymbol{n})/q_j^-(\boldsymbol{n}+\boldsymbol{e}_j)$ leading to the nine remarkable Pólya splitting distribution of Table 1 at equilibrium.

# 4 Discussion

Our main contribution is to connect the class of Polya-splitting distribution to the neutral theory of biodiversity in ecology, a useful null model allowing the evaluation of non-neutral processes such as adaptation or natural selection (Alonso et al., 2006). We found that for any Polya-splitting distribution, there exists a multivariate jump process of neutral species with such stationnary distribution. However, staying at the very general level for the sum distribution, the associated transition rates may not have a straightforward biological interpretation. We therefore exhibited, nine transition rates parametrization with meaningful biological interpretation leading to usual parametric distributions.

However, Pólya splitting distributions induce only two types of dependence structures: either all species are independent or fully dependent with homogenized correlation sign. To extend this binary setting towards more complex nested dependence structures between species or communities, we suggest the use of recursive application of splitting distributions along a partition tree of species.

Otherwise, the inclusion of environmental factors in Pólya splitting distributions is a natural extension. It could be achieved assuming a regression model for the sum distribution and another for the Pólya distribution (see Peyhardi et al. (2021) for more details in the multinomial splitting regression context).

Finally, combining a partition tree approach with the inclusion of environmental covariates at each node leads to propose nested multi-level inhomogenous splitting models. Such models should be interesting alternatives to classical approaches used in joint species distribution contexts mainly based on conditional Independence's (Warton et al., 2015; Ovaskainen and Abrego, 2020) and the use of the multivariate Poisson log-Normal distribution. Comparatively, our approach allows to model dependencies between species at the observation level, while the classical JSDM's model dependencies at the latent process strata.

# References


Alonso, D., Etienne, R.S., McKane, A.J., 2006. The merits of neutral theory. Trends in Ecology & Evolution 21, 451–457. doi:10.1016/j.tree.2006.03.019.

Bol'shev, L.N., 1965. On a characterization of the poisson distribution and its statistical applications. Theory of Probability & Its Applications 10, 446–456.

Etienne, R., Alonso, D., McKane, A., 2007. The zero-sum assumption in neutral biodiversity theory. Journal of theoretical biology. 248, 522–536.

Hubbell, S.P., et al., 2001. The unified neutral theory of biodiversity and biogeography. volume 32. Princeton University Press Princeton.

Magurran, A.E., 2021. Measuring biological diversity. Current Biology 31, R1174–R1177.

Ovaskainen, O., Abrego, N., 2020. Joint Species Distribution Modelling: With Applications in R. Ecology, Biodiversity and Conservation, Cambridge University Press. doi:10.1017/9781108591720.

Peyhardi, J., 2023. On quasi Pólya thinning operator. Brazilian Journal of Probability and Statistics 37, 643 – 666.

Peyhardi, J., Fernique, P., Durand, J.B., 2021. Splitting models for multivariate count data. Journal of Multivariate Analysis 181, 104677.

Peyhardi, J., Laroche, F., Mortier, F., 2024. Pólya-splitting distributions as stationary solutions of multivariate birth-death processes under extended neutral theory. Journal of Theoretical Biology .

Rao, B.R., Janardan, K., 1984. The use of the generalized markov-polya distribution as a random damage model and its identifiability. Sankhyā: The Indian Journal of Statistics, Series A , 458–462.

Rao, C.R., 1965. On discrete distributions arising out of methods of ascertainment. Sankhyā: The Indian Journal of Statistics, Series A , 311–324.

Warton, D., Blanchet, F., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S., Hui, F., 2015. So many variables: Joint modeling in community ecology. Trends in Ecology and Evolution 30, 766–779.

# Inférence de réseaux d'associations à l'échelle de groupes à partir de données d'abondance avec le modèle PLN-Block

Jeanne Tous [1] & Julien Chiquet [2]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, julien.chiquet@inrae.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, jeanne.tous@inrae.fr*

**Résumé.** Les réseaux d'association constituent un outil utile en écologie pour identifier des relations entre espèces qui ne peuvent être expliquées par les variables environnementales observées, et peuvent donc aider à la compréhension du fonctionnement de systèmes complexes. De tels réseaux peuvent être inférés à partir de données d'abondance, comme des comptages d'espèces en écologie, grâce à PLN-network [Chiquet, Mariadassou et Robin, 2019], une méthode d'inférence de réseaux fondée sur un modèle Poisson-log-normal couplée à une procédure d'estimation de type GLASSO [Friedman et al., 2007]. Cependant, lorsque le volume de données et le nombre d'espèces étudiés augmentent, les réseaux obtenus sont complexes à étudier puisque les associations entre espèces sont identifiées avec des degrés de confiance variés et leur interprétation individuelle est sujette à caution. Il existe des métriques pour agréger l'information contenue dans ces réseaux (nombre total d'associations, intensité moyenne de celles-ci, nombre de cliques dans le réseaux...). Cependant, ces métriques résument la complexité des réseaux à un faible nombre d'informations, nécessairement réductrices. Un compromis entre l'échelle très fine de chaque association et l'échelle globale de ces métriques peut être obtenu par un clustering sur les sommets du graphe et l'inférence d'un réseau à l'échelle des clusters plutôt que celle des espèces. De tels groupes rassembleraient des sommets dont les positionnements vis-à-vis du reste du réseau semblent similaires. Nous proposons le modèle PLN-Block, une méthode d'inférence de réseaux dérivée de PLN-network, qui vise à simultanément effectuer un clustering des espèces étudiées et inférer une structure de réseau entre les clusters, à partir de données d'abondance. Nous introduisons ici ce modèle, discutons les résultats du clustering et les améliorations qui peuvent lui être apportées, notamment pour mieux tenir compte de la nature spécifique des données écologiques.

**Mots-clés.** Modèles poisson-log-normaux ; clustering ; inférence variationnelle ; biodiversité ; réseaux d'associations.

**Abstract.** Association networks are a useful tool in fields such as ecology to identify relationships that cannot be explained by observed environmental factors and thus help understand the functioning of complex systems. Such networks can be inferred from abundance data, such as measured presence in ecology thanks to PLN-network [Chiquet, Mariadassou and Robin, 2019], a Poisson-log-normal based network inference method associated with an estimation procedure like GLASSO [Friedman et al., 2007]. However, as observed data and the number of studied species get more numerous, the resulting network becomes very tedious to study since pairwise associations can be identified with various degree of confidence and edges' individual interpretation can be questionable. Measures exist to summarize them thanks to a few metrics, such as the overall number of associations, their intensity, the number of cliques... but these metrics, on the other hand, reduce a very complex structure to a few simple figures, necessarily wanting. A compromise between the micro-scale of pairwise associations and the macro-scale of such measures could be to cluster graph vertices and infer a network at the scale of these clusters. Such groups would gather vertices that seem to behave similarly in regard to their positioning and associations profiles in the network. We introduce PLN-block, a network inference method derived from PLN-network that aims at simultaneously clustering studied species and derive a network structure between these clusters from abundance data. We introduce the model and the associated inference method, discuss the clustering results and how the model could be improved to better adapt to the specificity of ecological data.

**Keywords.** Poisson log-normal models, clustering, variational inference, biodiversity, association networks.

# 1 Introduction

Les réseaux d'association sont des objets mathématiques utiles pour décrire les relations entre espèces en écologie. Ils peuvent être inférés à partir de données d'abondance d'espèces : les relations de dépendance entre ces abondances sont mesurées par des outils de type corrélation et les dépendances qui ne peuvent être expliquées par les seules covariables sont identifiées comme des associations. Dans

certains cas, ces associations peuvent être vues comme des interactions, mais souvent elles n'en sont que des approximations, et l'interprétation de ces réseaux est plus limitée puisque les associations identifiées peuvent aussi être expliquées par des réactions similaires à des facteurs environnementaux non mesurés [Poggiato et. al, 2021].

L'inférence de réseaux menée dans notre cadre de travail se fonde sur des modèles graphiques non orientés [Lauritzen, 1996] ou champs aléatoires de Markov [Harris, 2016]. Dans ce cadre, les espèces $i$ et $j$ sont considérées comme liées dans le réseau si et seulement si leurs abondances sont conditionnellement dépendantes, au sens du modèle statistique défini, étant données les covariables et les abondances des autres espèces. Inférer un réseau d'association revient alors à calculer des corrélations partielles entre des variables liées aux abondances de chaque espèce, comme nous le détaillerons dans la description de PLN-block.

Du fait de la nature des données d'abondance, à savoir des données de comptage, le modèle statistique considéré ici doit être adapté aux données discrètes, ce qui n'est pas le cas des modèles graphiques gaussiens classiques. PLN-network [Chiquet, Mariadassou, Robin, 2019] est un modèle conçu pour tenir compte de la spécificité des données de comptage pour inférer des réseaux peu denses. Les données y sont modélisées par une distribution Poisson log-normale [Aitchison et Ho, 1989], tenant compte des covariables potentielles et intégrant une variable latente pour modéliser les dépendances résiduelles qui définissent le réseau. L'ajout d'une contrainte de sparsité sur le réseau reconstruit permet de sélectionner seulement les dépendances les plus fortes et de limiter la sélection de celles qui semblent les moins certaines.

Cependant, le réseau obtenu peut être complexe à interpréter, particulièrement lorsqu'un grand nombre d'espèces est étudié. En effet, il s'agit d'un objet complexe, informatiquement lourd à manipuler, et riche en informations. De plus, les associations résultantes doivent être considérées avec précaution dans la mesure où elles sont le résultat d'une procédure statistique et sont identifiées avec des degrés de confiance divers qui dépendent du modèle et de la sparsité du réseau. Le choix de l'hyper-paramètre qui définit le niveau de sparsité imposé dans le réseau permet ainsi de modifier le niveau de certitude avec lequel les associations sont retenues. Il existe des métriques pour agréger les informations contenues dans le réseau (nombre total d'associations, intensité moyenne de celles-ci...) mais elles semblent en revanche très réductrices au vu de la complexité des objets étudiés.

Une autre méthode consisterait à effectuer un clustering des nœuds du réseau, fondé sur leur profil d'associations avec les autres nœuds. Un tel résultat peut être

obtenu avec des Stochastic Block Models (SBM), introduits par Holland et al. en 1983 [Holland et al., 1983]. Dans leur formulation la plus simple, il s'agit de modéliser des associations binaires par un modèle de mélange : les associations suivent une loi de Bernouilli dont le paramètre est entièrement défini par les groupes auxquels appartiennent chacun des nœuds considérés. De nombreuses variations de ce modèle existent, par exemple pour tenir compte des distributions d'interactions possibles (gaussienne, Poisson...) de l'évolution de la composition des groupes [Yang et al., 2011], [Matias et Miele, 2017] ou des incertitudes sur les données observées [Rebafka et al., 2019]. Cependant, les SBM sont appliqués à des réseaux déjà reconstruits, de sorte que les résultats du clustering ne peuvent pas fournir d'information pour nourrir la construction initiale du réseau à l'échelle des clusters.

Avec PLN-Block, nous proposons d'effectuer le clustering des espèces observées et l'inférence d'un réseau d'associations à l'échelle de ces clusters simultanément grâce à une procédure d'optimisation alternée.PLN-block est directement dérivé de PLN-network par l'ajout d'une couche latente supplémentaire pour décrire les groupes auxquels les espèces appartiennent, dont découlent leurs probabilités d'association, dans la logique de modèle de mélange des SBM. Pour l'inférence des paramètres du modèle, nous utilisons une méthode d'inférence variationnelle, fondée sur le calcul d'une approximation de la fonction de vraisemblance.

Nous présentons le modèle PLN-block dans la section 2. La section 3 décrit la méthode d'inférence utilisée, tandis que la 4 présente les résultats de premières simulations. Enfin nous discutons des possibilités d'amélioration du modèle dans la section 5.

# 2 Modèle

Dans ce modèle,

- on considère $p$ espèces observées dans $n$ sites, et l'on suppose qu'elles se divisent en $Q$ groupes ($Q$ est un hyper-paramètre) ;
- la variable $Y_{i,j}$ décrit l'abondance (le nombre d'individus observés) de l'espèce $j$ dans le site $i$ ;
- $X_i \in \mathbb{R}^m$ est le vecteur des covariables du site $i$ et $B_j$ le vecteur de paramètres de régression correspondant pour l'espèce $j$, supposé indépendant du site con-

sidéré. On note $X \in \mathcal{M}_{n,m}(\mathbb{R})$ la matrice dont la $i$-ème ligne est $X_i$, et $B \in \mathcal{M}_{m,p}(\mathbb{R})$ la matrice dont la $j$-ème colonne est $B_j$ ;

- la variable latente $Z_i \in \mathbb{R}^Q$ décrit la structure de dépendance entre les groupes d'espèces sur le site $i$ : pour tout $i, Z_i \sim \mathcal{N}(0, \Sigma)$ ;
- la variable latente $C_j$ indique le groupe auquel l'espèce $j$ appartient. Pour tout $j, C_j \sim \mathcal{M}(1, \alpha)$ avec $\alpha = (\alpha_q)_{1 \leq q \leq Q}$ et $\sum_{q=1}^{Q} \alpha_k = 1$ ; $\alpha_q$ est donc la probabilité pour une espèce d'appartenir au groupe $q$ ; on note $\alpha = (\alpha_q)_{1 \leq q \leq Q}$ et $C_{j,q} = \mathbb{1}_{C_j = q}$, $C_{j,q}$ vaut 1 si l'espèce $j$ appartient au groupe $q$, 0 sinon ;
- $o_{i,j}$ décrit un offset pour l'espèce $j$ sur le site $i$, pour permettre, le cas échéant, de tenir compte des différences d'efforts d'échantillonnage entre sites / espèces.

On suppose que chaque $Y_{i,j}$ suit, conditionnellement à $Z_i$ et $C_j$ une distribution Poisson log-normale dont le paramètre est déterminé par les covariables et par le $Z_{i,q}$ correspondant au groupe $q$ auquel appartient l'espèce $j$ :

$$Y_{i,j} | Z_i, C_j \sim \mathcal{P}(\exp(x_i^T B_j + o_{i,j} + \textstyle\sum_{q=1}^{Q} Z_{i,q} C_{j,q})).$$

Le réseau entre les groupes est donc traduit par la variable latente $Z$ ou, plus exactement, par sa matrice de précision $\Omega = \Sigma^{-1}$. En effet, la corrélation partielle entre $Z_{q_1}$ et $Z_{q_2}$ est donnée par $\rho_{q_1,q_2} = \frac{-\Omega_{q_1,q_2}}{\sqrt{\Omega_{q_1,q_1}\Omega_{q_2,q_2}}}$. Enfin, on peut ajouter une contrainte de sparsité sur $\Omega$ afin de contrôler la densité du réseau. Pour ce faire, on ajoute une pénalité $\ell_1$ sur les termes non diagonaux de $\Omega$, multipliée par un hyper-paramètres $\lambda$ qui *in fine* contrôle le nombre de corrélations partielles non nulles, donc le nombre d'arêtes du réseau inféré.

# 3 Méthode d'inférence

On note $\theta = (B, \Omega, \alpha)$ les paramètres du modèle. Il n'existe pas de forme explicite de la vraisemblance complète du modèle, $\sum_C \int_{-\infty}^{+\infty} p_\theta(Y, Z, C) dZ$ et il n'est donc pas possible de calculer directement $\theta$ par maximum de vraisemblance. Une simple stratégie EM [Dempster et al., 1977] ne résout pas non plus le problème car il n'existe pas de formule explicite pour calculer les moments de $p_\theta(Z, C|Y)$.

On opte donc pour une stratégie d'inférence variationnelle. Pour approcher la distribution conditionnelle $p_\theta(Z, C|Y)$ on fait une approximation de champ moyen

[Blei et al., 2017]. On cherche donc $\pi_\psi$ dans $\Pi = \{\pi_{\psi_1}(Z)\pi_{\psi_2}(C)\}$ ($\psi = \{\psi_1, \psi_2\}$ désigne l'ensemble des paramètres variationnels) pour approcher $p_\theta(Z, C|Y)$, avec :

- $\pi_{\psi_1}$ : pour tout $i, \pi_{\psi_1}(Z_i) \sim \mathcal{N}(M_i, S_i)$, avec $S_i = diag(s^2_{i,q})_{q\in[\![1;Q]\!]}$ diagonale pour tout i. On note $S \in \mathcal{M}_{n,K}(\mathbb{R})$ définie par $S_{i,q} = s^2_{i,q}$ et $M \in \mathcal{M}_{n,K}(\mathbb{R})$ définie par $M_{i,q} = M_{i_q}$ de sorte que $\psi_1 = (M, S)$.

- $\pi_{\psi_2}$ : pour tout $j, C_j \sim \mathcal{M}(1, (\tau_{q,j})_{1\leq q\leq Q})$ avec pour tout $j, \sum_{q=1}^{Q} \tau_{q,j} = 1$. On note $\tau \in \mathcal{M}_{Q,p}(\mathbb{R})$ la matrice définie par $(\tau_{q,j})_{j\in[\![1;p]\!],q\in[\![1;Q]\!]}$ de sorte que $\psi_2 = (\tau)$.

On fait également l'hypothèse que, sous $\pi_\psi$ les $Z_i$ sont independants entre eux, de même que les $C_j$. La tâche de clustering revient donc à trouver $\max((\tau_{q,j})_{1\leq q\leq Q})$ pour tout $1 \leq j \leq p$.

Dans le cadre de l'approximation variationnelle, on peut définir une vraisemblance approximative (ELBO) à maximiser :

$$\begin{aligned} J(Y;\theta,\psi) &= \log(p_\theta(Y)) - \mathrm{KL}(\pi_\psi(Z,C)||p_\theta(Z,C|Y)) \\ &= \mathbb{E}_\pi(p_\theta(Y,Z,C)) - \mathbb{E}_\pi(\log(\pi_{\psi_1}(Z))) - \mathbb{E}_\pi(\log(\pi_{\psi_2}(C))) \end{aligned}$$

Les calculs permettent d'aboutir à une expression matricielle explicite pour $J$.

Dans le cas où on ajoute une contrainte de sparsité sur $\Omega$, on peut définir une nouvelle fonction objectif : $J_2 = J - \lambda\|\Omega\|_{\ell_1,\text{off}}$ où $\lambda\|\Omega\|_{\ell_1,\text{off}}$ désigne la norme de $\Omega$ considérée sans ses termes diagonaux (qui traduisent les variances au sein des groupes mais ne donnent pas d'informations sur les corrélations entre ces derniers), et $\lambda$ est un hyper-paramètre.

La procédure d'inférence consiste ensuite à mettre alternativement à jour $B, M, S$, $\Omega$ et $\tau, \alpha$ pour maximiser $J$ ou $J2$, dans un cadre d'EM variationnel que l'on ne détaille pas ici.

# 4 Simulations

Les simulations effectuées jusqu'à présent visent d'abord à tester les capacités de clustering de l'algorithme d'inférence. Nous avons donc simulé des données sous le modèle décrit en section 2 et évaluer les capacités de l'algorithme à identifier

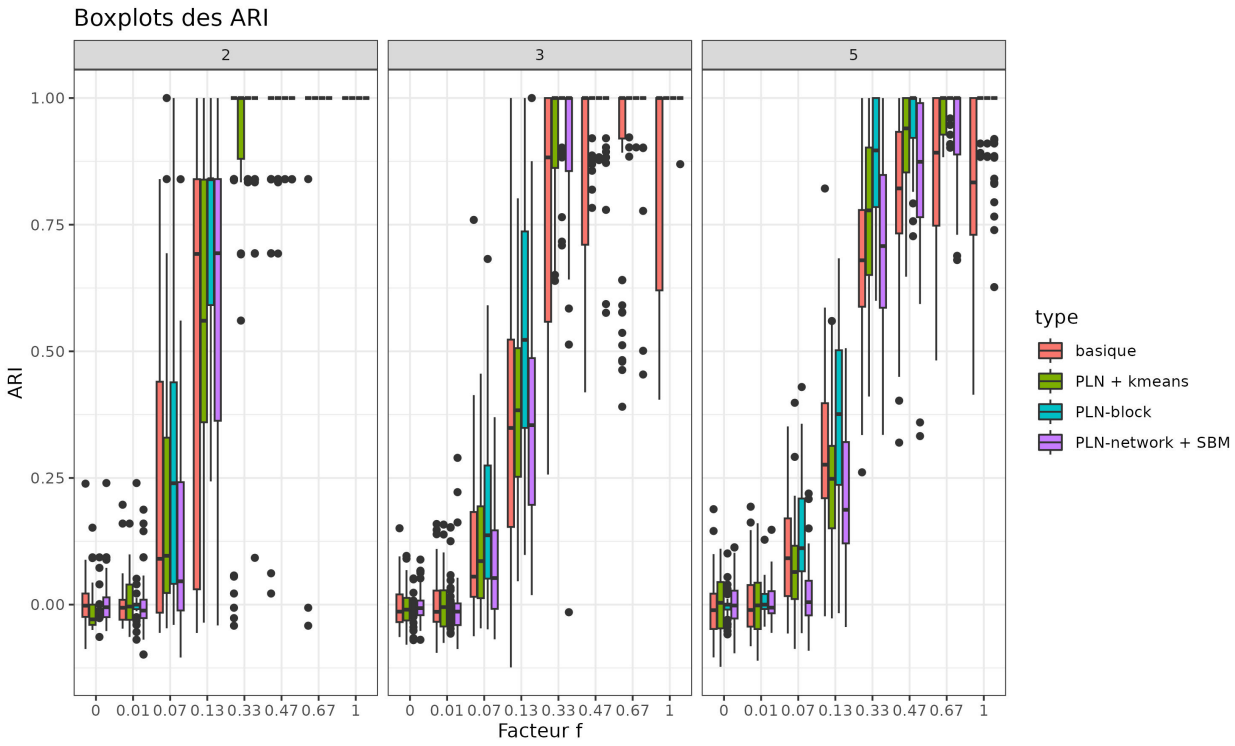



Figure 1: ARI des clusterings effectués sur des données simulées sous PLN-block avec $p = 25$, $n = 50$, $q = 2$, 3 ou 5. Abscisse : facteur par lequel est multipliée la matrice $\Sigma$ réputée "facile".

correctement ces clusters par l'average rand index (ARI) entre les vrais clusters et les clusters inférés.

Nous avons comparé les ARI obtenus avec notre algorithme à ceux obtenus par différentes méthodes :

1. Une méthode "basique" qui consiste à effectuer une régression de Poisson sur les données suivi d'un clustering k-means sur les colonnes de résidus.
2. Une méthode "PLN + k-means" qui consiste à appliquer un modèle PLN simple aux données suivi d'un clustering k-means sur les résidus, obtenus grâce aux moyennes variationnelles de la variable latente $Z$.
3. Une méthode "PLN-network + SBM" : on reconstruit un réseau à l'échelle des

espèces avec PLN-network puis on lui applique un modèle SBM pour en tirer des clusters.

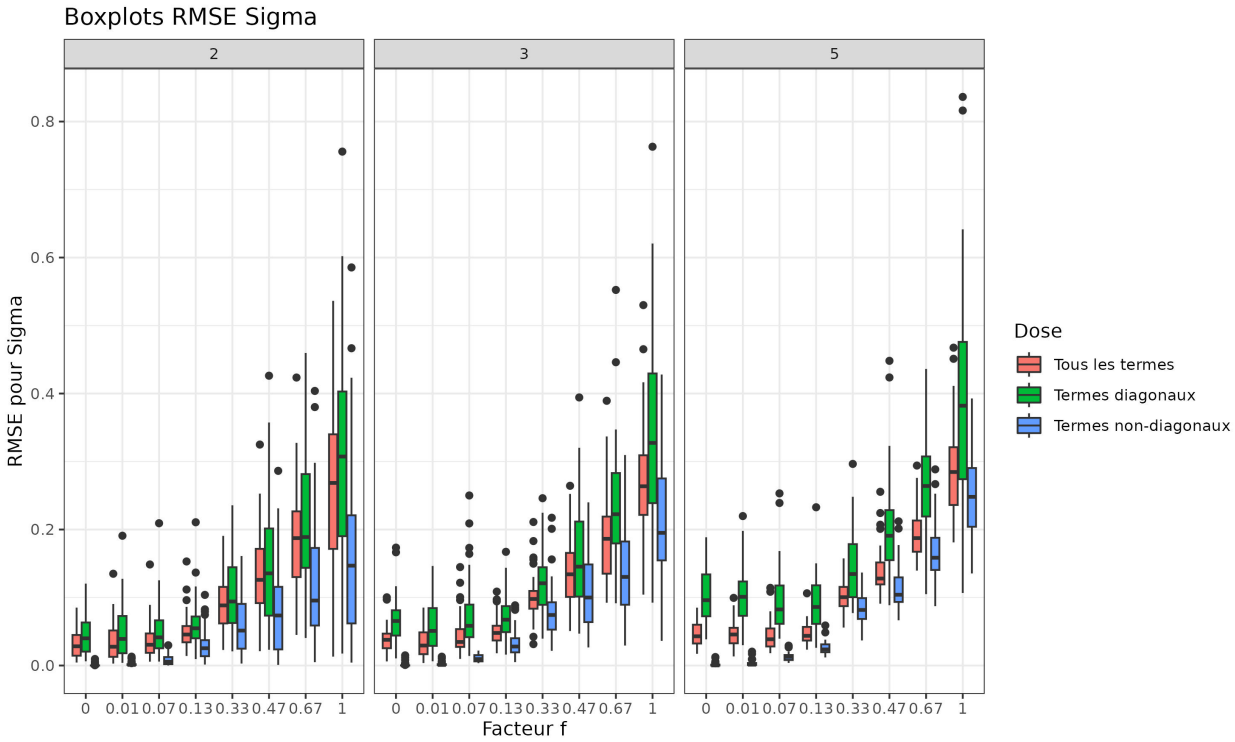



Figure 2: RMSE de $\Sigma$ sur des données simulées sous PLN-block avec $p = 25$, $n = 50$, $q = 2$, 3 ou 5. Abscisse : facteur par lequel est multipliée la matrice $\Sigma$ initiale.

Nous proposons ici de premières simulations avec $p = 25$ espèces, $n = 50$ sites. On tire aléatoirement les covariables environnementales (en deux dimensions) contenues dans $X$ entre 0 et 1, et les paramètres de régression associés, contenus dans $B$, entre -1 et 1. Nous effectuons les tests pour $q = 2, 3$ ou 5 groupes. Pour chaque $q$, on a créé à la main une matrice définie positive $\Sigma$ par laquelle on a cherché à mettre des contrastes importants entre les corrélations afin de rendre le clustering plus facile. Ensuite, nous avons simulé avec $\Sigma$ multipliée par un facteur $f$ entre 0 et 1. L'objectif est de rendre la tâche de clustering plus difficile lorsque $f$ se rapproche de 0 car l'effet des groupes sur la variable $Y$ observée est alors moins fort par rapport à celui des covariables environnementales. Les données sont simulées sous le modèle PLN-block.

La figure 1 montre les résultats de cette simulation. On constate que la tâche de clustering se complexifie effectivement beaucoup lorsque $f$ se rapproche de 0, et ce quelle que soit la méthode employée. La méthode PLN-block semble plus robuste que les autres dès que $f$ dépasse 0.07, constat plus marqué pour $q = 5$. L'algorithme d'optimisation semble donc aboutir aux résultats souhaités pour le clustering, dès lors que l'effet des groupes est suffisamment important. La figure 2 montre en revanche une augmentation de la RMSE pour l'estimation de $\Sigma$ à mesure que $f$ augmente, ce qui pourrait s'expliquer par une plus grande variance dans l'estimation de plus grandes valeurs de $\Sigma$.

Nous présenterons oralement des résultats concernant les réseaux inférés par PLN-block et les effets de l'hyper-paramètre de sparsité.

# 5 Discussion

PLN-block propose un modèle d'inférence de réseaux à l'échelle de groupes identifiés par ce même modèle. Les premiers résultats sur la tâche de clustering montrent la capacité du modèle à inférer correctement les clusters dès lors que leur effet sur les observations dépasse un certain seuil, qui semble dépendre du nombre de clusters. Cependant la robustesse du modèle dans des cas plus réalistes, pour lesquels des effets spécifiques des points du graphes s'additionnent à ceux des groupes n'est pas assurée. En particulier, il semblerait judicieux d'intégrer dans le modèle un effet spécifique des points (en plus de l'effet des clusters) sur la variance. Sans cela, , il se peut que des points soient regroupés dans le même cluster en raison de variances similaires (diagonale de $\Omega$) et non en raison de corrélations similaires, ce qui est la visée de PLN-block. Nous travaillons actuellement à cette généralisation.

Dans un second temps, d'autres développements de PLN-block pourraient intégrer une dimension temporelle pour tenir compte de l'évolution des groupes ou des associations entre eux, ou encore d'y ajouter des *a priori* écologiques pour guider les tâches de clustering et d'inférence de réseau.

# Bibliographie

Aitchison, J., et Ho, C. H. (1989), The multivariate Poisson-log normal distribution, *Biometrika*, 76, pp 643-653.

Blei, D. M., Kucukelbir, A., et McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Chiquet, J. Mariadassou, M. et Robin, S. (2019) Variational Inference of Sparse Network from Count Data, *Proceedings of Machine Learning Research*, 97, pp 1162-1171.

Chiquet, J. Mariadassou, M. et Robin, S. (2021) The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances., *Frontiers in ecology and evolution*, 9, p 588292.

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1), 1-22.

Friedman, J., Hastie, T. et Tibshirani, R. (2007) Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, 9(3), pp 432-441.

Harris, D.J. (2016) Inferring species interactions from co-occurrence data with Markov networks, *Ecology*, 97, pp 3308-3314.

Holland, P. W., Laskey, K. B. et Leinhardt, S.(1983) Stochastic block models: First steps, *Social networks*, 5, pp 109-137.

Lau, M. K., Borrett, S. R., Baiser, B., Gotelli, N. J. et Ellison, A. M.(2017) Ecological network metrics: opportunities for synthesis, *Ecosphere*, 8.

Lauritzen,S.L. (1996) *Graphical Models*, 17.

Matias, C. et Miele, V.(2017) Statistical clustering of temporal networks through a dynamic stochastic block model, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4), pp 1119-1141.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. et Thuiller, W. (2021) On the Interpretations of Joint Modeling in Community Ecology, *Trends in Ecology and Evolution*, 2806.

Rebafka, T., Roquain, E. et Villers, F. (2019) On the Graph inference with clustering and false discovery rate control *rXiv preprint arXiv:1907.10176*.

Yang, T., Chi, Y., Zhu, S., Gong, Y. et Jin, R.(2011) Detecting communities and their evolutions in dynamic social networks—a Bayesian approach, *Machine learning*, 82, pp 157-189.

# Zero-inflation in the MPLN Family

Bastien Batardière [1], Julien Chiquet [1], François Gindraud [2]& Mahendra Mariadassou [3]

[1] *MIA Paris-Saclay, Université Paris-Saclay, AgroParisTech, INRAE, France, {prenom.nom}@inrae.fr*
[2] *Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, francois.gindraud@inria.fr*
[3] *Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France, mahendra.mariadassou@inrae.fr*

**Résumé.** Les données de comptage en haute dimension sont difficiles à analyser telles quelles, et les approches basées sur des modèles statistiques restent efficaces et appropriées tout en préservant l'interprétabilité. Le modèle Poisson-Log-Normal (PLN) (multivarié) en est un exemple et suppose que les données de comptage sont influencées par une variable gaussienne latente structurée, exploitant les dépendances entre les comptages. Bien que les données de comptage du monde réel soient fréquemment caractérisées par des excès de zéros, un tel modèle ne prend pas en compte cette réalité. Nous proposons ici le modèle Zero-Inflated PLN (ZI-PLN), ajoutant une composante multivariée à excès de zéros au modèle, sous la forme d'une variable latente Bernoulli supplémentaire. L'inflation de zéros peut être fixe, spécifique au site, spécifique à la variable ou dépendre de covariables. Nous estimons les paramètres du modèle en utilisant une inférence variationnelle et comparons deux approximations : (i) distributions variationnelles gaussiennes et bernoulli indépendantes ou (ii) distribution gaussienne variationnelle conditionnée à la distribution bernoulli. La méthode est évaluée sur des données synthétiques. Tous les algorithmes sont disponibles dans un package Python `pyPLNmodels` et un package R `PLNmodels`.

**Mots-clés.** Données de comptages, modèle Poisson Log Normal, zéro-inflation, inférence variationnelle, optimisation alternée.

**Abstract.** High-dimensional count data are hard to analyze as is and statistical model-based approaches remain efficient and adequate, while preserving explainability. The (multivariate) Poisson-Log-Normal (PLN) model is one of them and assumes count data are driven by an underlying structured latent Gaussian variable, exploiting dependencies between counts. While real-world count data are frequently zero-inflated, such a model does not account for this reality. Here we propose the Zero-Inflated PLN (ZI-PLN) model, adding a multivariate zero-inflated component to the model, as an additional Bernoulli latent variable. The Zero-Inflation can be fixed, site-specific, feature-specific or depend on covariates. We estimate model parameters using variational inference and compare two approximations: (i) independent Gaussian and Bernoulli variational distributions or (ii) Gaussian variational conditioned on the Bernoulli one. The method is assessed on synthetic data. All the algorithms are available in a python package `pyPLNmodels` and an R package `PLNmodels`

**Keywords.** Count data, Poisson lognormal model, zero inflated model, variational Inference, alternate optimisation

# 1 Introduction

Count data appears in many different fields such as ecology, accidents analysis, single-cell RNA (scRNA) sequencing and metagenomics. For example, researchers may be interested in estimating the correlation between abundances of different species or expression of different genes in a cell. More specifically, the model introduced in this paper is motivated by the increasing importance of microbiome studies. Broadly speaking, a microbiome is a collection of microbes, together with their genomes, found in a given habitat (*e.g.* plant leaves, human gut, waste water, etc.). The most widespread way of studying microbiomes is to amplify and sequence a marker gene, which acts as a molecular barcode. The sequences are processed through bioinformatics pipelines [Escudié et al., 2017] to produce Operational Taxonomic Units (OTUs) / Amplicon Sequence Variants (ASVs), a proxy for microbial species in microbial ecology, and enumerated to create count tables, recording the abundance of each OTU/ASV in each sample. Those tables are characterized by a very high fraction (ranging from 80 to 95%) of zero counts and a high number of variables.

The (multivariate) Poisson-Log-Normal [in short PLN, see Aitchison and Ho, 1989] model offers a general framework to multivariate count data by offering flexibility to describe dependencies between counts by means of a latent Gaussian variable. By means of Poisson emission law with Log-Normal parameter, PLN models results in a overdispersed models. The underlying Gaussian structure inherent to the PLN model makes correlation between variables natural, unlike its NB counterpart. More generally, the PLN model falls in the family of latent variable models (LVMs), and more specifically of multivariate generalized linear mixed models (mGLMMs) sometimes also called generalized linear latent variable models (GLLVMs). In those models, the distribution of observed responses usually belongs either to the exponential family (Bernoulli, Binomial, Poisson, Negative-Binomial, with or without Zero-Inflation, etc.) or the exponential dispersion model (Tweedie, etc.). Model parameters are related to linear combinations of latent variables (and possibly covariates) through a simple link function. Parameter estimation for common GLLVMs is efficiently implemented in some packages [Niku et al., 2019, Seabold and Perktold, 2010], making a popular option for multivariate count data. However, while some models allows for dependency between variables in the latent space and other accounts for zero-inflation, no model accounts, to the best of our knowledge, for both at the same time.

We introduce here the Zero-Inflated Poisson Log-Normal (ZI-PLN) model, based on the Poisson Log-Normal (PLN) model. ZI-PLN benefits from the Gaussian structure of the PLN model, with an extra zero-inflated component. This extra layer adds flexibility to the model as its parameters can be chosen to be shared across the individuals, across the genes or even to depend on its own set of covariates. As exact inference of (ZI)PLN is intractable and conditional laws are only partially known, we cannot rely on the Expectation-Maximization (EM) algorithm [Dempster et al., 1977], as done for optimizing classical latent variable models. We rely on variational inference [Jaakkola and Jordan, 2000, Wainwright and Jordan, 2008, Hui et al., 2017, Blei et al., 2017]. Other approaches based on Monte Carlo techniques have been proposed [JAC, 2007, Cap] to infer the maximum likelihood estimator, but it does not scale with the dimension of the observations. Numerical integration can be performed [Aitchison and Ho, 1989] as an alternative to the variational approximation used

here but becomes prohibitive when the number of dimensions exceeds 5. Here, we develop a Variational-EM algorithm where we propose two different variational approximations. The first assumes conditional independence between both components, resulting in a fast M step. By contrast, the second is slightly slower but leverages the dependence between components to use a more complex variational approximation.

**Related work** ZINBWaVE, proposed by Risso et al. [2018] is the closest work to ours, modelling zero-inflation (resp. counts) as a logistic (resp. log-linear) regression involving sample-level, gene-level and (unobserved) sample-level covariates, where the unobserved covariates are presumed to be unwanted variations and captured through latent factors. This model however suffers from a lack of identifiability and is mostly interested in estimating the probability that a null count arises from zero-inflation. We distinguish ourselves from ZINBWaVE via identifiability of parameters and most importantly via the inherent and explicit dependency structure between variables.

# 2 Model

**Background: Multivariate Poisson lognormal-model** The multivariate Poisson lognormal model relates a $p$-dimensional observation count vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip}) \in \mathbb{N}^p$ to a $p$-dimensional vector of Gaussian latent variables $\mathbf{Z}_i \in \mathbb{N}^p$ with precision matrix $\mathbf{\Omega}$ (that is, covariance matrix $\mathbf{\Sigma} \triangleq \mathbf{\Omega}^{-1}$). We adopt a formulation of PLN close to a multivariate generalized linear model, where the main effect is due to a linear combination of $d$ covariates $\mathbf{x}_i \in \mathbb{R}^d$ (including a vector of intercepts). We also let the possibility to add some offsets for the $p$ variables in in each sample, that is $\boldsymbol{o}_i \in \mathbb{R}^p$:

$$\begin{aligned} \text{latent space} \quad & \mathbf{Z}_i \sim \mathcal{N}(\mathbf{o}_i + \mathbf{x}_i^\intercal \boldsymbol{B}, \mathbf{\Omega}^{-1}), \\ \text{observation space} \quad & Y_{ij} \,|\, Z_{ij} \quad \text{indep.} \qquad \boldsymbol{Y}_i | \boldsymbol{Z}_i \sim \mathcal{P}\left(\exp\{\boldsymbol{Z}_i\}\right). \end{aligned} \tag{1}$$

The $d \times p$ matrix $\boldsymbol{B}$ is the latent matrix of regression parameters. The latent covariance matrix $\mathbf{\Sigma}$ describes the underlying residual structure of dependence between the $p$ variables, once the covariates are accounted for. We denote by $\boldsymbol{Y}, \boldsymbol{O}, \boldsymbol{X}$ the observed matrices with respective sizes $n \times p, n \times p$ and $n \times d$ stacking row-wise the vectors of counts, offsets and covariates (respectively $\mathbf{Y}_i, \mathbf{x}_i$ and $\mathbf{o}_i$). We also denote by $\boldsymbol{Z}$ the $n \times p$ matrix of unobserved latent Gaussian vectors $\mathbf{Z}_i$.

**Zero-inflated PLN regression model** We now aim to model an excess of zeros in the data by adding zero-inflation to the standard PLN model eq. (1), so that the zeros in $\boldsymbol{Y}_i$ arise from two different sources: either from a component where zero is the only possible value, or from a standard PLN component like eq. (1). This two-component mixture is described thanks to an additional latent vector $\boldsymbol{W}_i = (W_{i1,\ldots,ip}) \in \mathbb{R}^p$ of Bernoulli random variables, parametrized by probabilities $\pi_i = (\pi_{i1}, \ldots, \pi_{ip})$ describing the probability that variable $j$ in

sample $i$ belongs to the pure zero component:

$$\begin{aligned} \text{PLN latent space} \quad & \boldsymbol{Z}_i = (Z_{ij})_{j=1\ldots p} && \sim \mathcal{N}(\boldsymbol{o}_i + \mathbf{x}_i^\intercal \boldsymbol{B}, \boldsymbol{\Omega}^{-1}), \\ \text{excess of zero} \quad & \boldsymbol{W}_i = (W_{ij})_{j=1\ldots p} && \sim \otimes_{j=1}^{p} \mathcal{B}(\pi_{ij}), \\ \text{observation space} \quad & Y_{ij} \,|\, W_{ij}, Z_{ij} && \sim^{\text{indep}} W_{ij}\delta_0 + (1 - W_{ij})\mathcal{P}\left(\exp\{o_{ij} + Z_{ij}\}\right), \end{aligned} \tag{2}$$

where $\delta_0$ is the Dirac distribution and we note $\pi = (\pi_{ij})_{i=1..n, j=1..p}$. We focus on the probability $\pi_{ij}$ of belonging to the pure zero component. Our model is flexible enough to accomodate different parametrizations for $\pi_{ij}$ based on the availability of covariates and/or modeling choices made by the user:

$$\pi_{ij} = \pi \in [0, 1] \qquad \text{(non-dependent – ND)} \tag{3a}$$

$$\pi_{ij} = \text{logit}^{-1}(\boldsymbol{X}^0 \boldsymbol{B}^0)_{ij}, \;\; \boldsymbol{X}^0 \in \mathbb{R}^{n \times d_0}, \;\; \boldsymbol{B}^0 \in \mathbb{R}^{d_0 \times p} \qquad \text{(column-wise dependence – CD)} \tag{3b}$$

$$\pi_{ij} = \text{logit}^{-1}(\bar{\boldsymbol{B}}^0 \overline{\boldsymbol{X}}^0)_{ij}, \;\; \bar{\boldsymbol{B}}^0 \in \mathbb{R}^{n \times d_0}, \;\; \overline{\boldsymbol{X}}^0 \in \mathbb{R}^{d_0 \times p} \qquad \text{(row-wise dependence – RD)} \tag{3c}$$

where $\text{logit}^{-1}(\cdot)$ is the logistic (or inverse logit) function, $d_0 \geqslant 1$, $\boldsymbol{B}^0$ (resp. $\bar{\boldsymbol{B}}^0$) are regression coefficients associated with row-wise matrix of covariates $\boldsymbol{X}^0$ (resp. column-wise covariates $\bar{\boldsymbol{X}}^0$), obtained by stacking the vectors $((\mathbf{x}_1^0)^\intercal, \ldots, (\mathbf{x}_n^0)^\intercal)$, which may or may not be the same as in $\boldsymbol{X}$, the matrix of covariates in the PLN component.

Using standard results on Poisson and Gaussian distribution, we easily derive the expectation and variance of the ZI-PLN regression model. Letting $A_{ij} \triangleq \exp\left(o_{ij} + \mu_{ij} + \sigma_{jj}/2\right)$ with $mu_{ij} = \mathbf{x}_i^\intercal \boldsymbol{B}_j$, then

$$\begin{aligned} \mathbb{E}\left(Y_{ij}\right) &= (1 - \pi_{ij}) A_{ij} > 0, \\ \mathbb{V}\left(Y_{ij}\right) &= (1 - \pi_{ij}) A_{ij} + (1 - \pi_{ij}) A_{ij}^2 \left(e^{\sigma_{jj}} - (1 - \pi_{ij})\right). \end{aligned}$$

In the following, we are interested in inferring the vector of parameters $\theta = (\boldsymbol{\Omega}, \boldsymbol{B}, \pi) \in \mathbb{S}_p^{++} \times \mathcal{M}_{p,d}(\mathbb{R}) \times \mathcal{M}_{n,p}([0, 1])$ where $\mathbb{S}_p^{++}$ is the set of $p \times p$ positive-definite matrices and $\mathcal{M}_{p,d}(\mathbb{R})$ the set of $p \times d$ real-valued matrices. We first show that model 2 is identifiable.

**Identifiability of ZI-PLN models** Identifiability results are available for the ZI Poisson model [Li, 2012] and can be generalized to the ZI-PLN regression model. To this end, we first consider the simple ZI-PLN model, (*i.e.* a ZI-PLN model without covariate and a single parameter $\pi$), with a single sample, in order to drop the index $i$:

$$\begin{aligned} \mathbf{W} = (W_j)_{j=1\ldots p} &\sim \mathcal{B}^{\otimes}(\pi) = (\pi_1) \otimes \cdots \otimes (\pi_p) \\ \mathbf{Z} = (Z_j)_{j=1\ldots p} &\sim \mathcal{N}_p(\mu, \boldsymbol{\Omega}^{-1}) \\ Y_j | W_j, Z_j &\sim W_j \delta_0 + (1 - W_j)\mathcal{P}(e^{Z_j}), \quad Y_j \perp Y_k | \mathbf{W}, \mathbf{Z} \end{aligned} \tag{4}$$

**Proposition 1.** *The simple ZI-PLN model defined in 4 with parameter $\theta = (\pi, \mu, \boldsymbol{\Omega})$ and parameter space $(0, 1)^p \times \mathbb{R}^p \times \mathbb{S}_p^{++}$ is identifiable.*

**Proposition 2.** *The ZI-PLN regression model 2 with zero-inflation defined as in eq. (3b) and parameter $\theta = (\boldsymbol{\Omega}, \boldsymbol{B}, \boldsymbol{B}^0)$ and parameter space $\mathbb{S}_p^{++} \times \mathcal{M}_{p,d}(\mathbb{R}) \times \mathcal{M}_{p,d}(\mathbb{R})$ is identifiable if and only if the $n \times d$ matrix of covariates $\mathbf{X}^0$ is full rank.*

# 3 Estimation by Variational Inference

Our goal is to maximize the marginal likelihood. In the framework of latent models, a standard approach (e.g. with *Expectation-Maximization* algorithms) uses the following decomposition by integrating over the latent variables $\mathbf{W}, \mathbf{Z}$

$$\log p_\theta(\mathbf{Y}) = \log \frac{p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})}{p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})} = \int_{\mathbf{W},\mathbf{Z}} \log \frac{p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})}{p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})} p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y}) d\mathbf{W} d\mathbf{Z}. \tag{5}$$

However, for the ZI-PLN model, it is untractable since the conditional distribution $p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})$ has no closed-form. To overcome this issue, we rely on a variational approximation of this distribution which will yield a lower bound of $\log p_\theta(\cdot)$ to be optimized: for observation $i$, we denote by $\widetilde{p}_\psi(\mathbf{Z}_i, \mathbf{W}_i)$ the approximation of $p_\theta(\mathbf{Z}_i, \mathbf{W}_i|\mathbf{Y}_i)$ where $\psi$ is a set of variational parameters to be optimized. Subtracting to the untractable expression eq. (5) of the log-likelihood the positive (and also untractable) quantity (known as the Kullback-Lebler divergence)

$$KL(\widetilde{p}_\psi(.) \| p_\theta(.|\mathbf{Y})) = \int_{\mathbf{W},\mathbf{Z}} \log \frac{\widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})}{p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})} \widetilde{p}_\psi(\mathbf{Z}, \mathbf{W}) d\mathbf{W} d\mathbf{Z}$$

results after some rearrangements in the following Evidence Lower BOund (ELBO):

$$\begin{aligned} J(\theta, \psi) &= \log p_\theta(\mathbf{Y}) - KL(\widetilde{p}_\psi(.) \| p_\theta(.|\mathbf{Y})) \\ &= \int_{\mathbf{W},\mathbf{Z}} \log \frac{p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})}{\widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})} \widetilde{p}_\psi(\mathbf{Z}, \mathbf{W}) d\mathbf{W} d\mathbf{Z} \\ &= \widetilde{\mathbb{E}}[\log p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})] - \widetilde{\mathbb{E}}[\log \widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})], \end{aligned} \tag{6}$$

which also looks like a plugin of integral eq. (5) with $p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})$ replaced with $\widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})$. An appropriate choice of variational approximation will make the integral calculation tractable, while leading to an acceptable approximation of the log-likelihood [Blei et al., 2017].

## 3.1 Choice of the variational family

**Standard variational approximation** A straightforward, yet efficient, approach is to consider the mean field approximation, which breaks all dependencies between the vectors $\mathbf{Z}_i$ and $\mathbf{W}_i$ and their respective coordinates and approximates the conditional distribution as the product of its coordinate-wise marginals:

$$\widetilde{p}^{(1)}_{\psi_i}(\mathbf{Z}_i, \mathbf{W}_i) \triangleq \widetilde{p}_{\psi_i}(\mathbf{Z}_i) \widetilde{p}_{\psi_i}(\mathbf{W}_i) = \otimes_{j=1}^p \widetilde{p}_{\psi_i}(\mathbf{Z}_{ij}) \widetilde{p}_{\psi_i}(\mathbf{W}_{ij}).$$

On top of that, we assume Gaussian and Bernoulli distribution for $\widetilde{p}_{\psi_i}(\mathbf{Z}_{ij})$ and $\widetilde{p}_{\psi_i}(\mathbf{W}_{ij})$ respectively, giving rise to the following variational approximation

$$\widetilde{p}^{(1)}_{\psi_i}(\boldsymbol{Z}_i, \boldsymbol{W}_i) = \otimes_{j=1}^p \mathcal{N}\left(M_{ij}, S_{ij}^2\right) \mathcal{B}\left(P_{ij}\right) \tag{7}$$

with $0 \leqslant P_{ij} \leqslant 1$ and $\psi_i = (M_{ij}, S_{ij}, P_{ij})_{1 \leqslant j \leqslant p}$. We denote $\boldsymbol{M}, \boldsymbol{S}$ and $\mathbf{P}$ the $n \times p$ matrices with respective entries $M_{ij}, S_{ij}$ and $P_{ij}$ $(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p)$. This approximation therefore requires the estimation of $3np$ additional variational parameters on top of $\theta$.

**Enhanced variational approximation.** As $W_{ij}$ can take only two values, the dependence between $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$ can easily be highlighted by noting that

$$Z_{ij}|W_{ij}, Y_{ij} = (Z_{ij}|Y_{ij}, W_{ij} = 1)^{W_{ij}} (Z_{ij}|Y_{ij}, W_{ij} = 0)^{1-W_{ij}} . \tag{8}$$

The conditional distribution of $Z_{ij}|Y_{ij}, W_{ij} = 1$ simplifies to $Z_{ij}|W_{ij} = 1$ and is thus known as $Z_{ij}$ and $W_{ij}$ are independent: it follows a gaussian distribution with mean $\boldsymbol{X}_i^\top B_j$ and variance $\Sigma_{jj}$. By contrast, $Z_{ij}|Y_{ij}, W_{ij} = 0$ is untractable and approximated by a gaussian distribution, giving rise to an alternative and slightly more involved variational approximation:

$$\widetilde{p}_{\psi_i}^{(2)}(\boldsymbol{Z}_i, \boldsymbol{W}_i) = \otimes_{j=1}^{p} \mathcal{N}(\boldsymbol{X}_i^\top B_j, \Sigma_{jj})^{W_{ij}} \mathcal{N}(M_{ij}, S_{ij}^2)^{1-W_{ij}} W_{ij}, \quad W_{ij} \sim^{\text{indep}} \mathcal{B}(P_{ij}) . \tag{9}$$

**Expected lower bounds** We set $\psi = (\psi_i)_{1 \leqslant i \leqslant n}$ and variational distribution $\widetilde{p}_{\psi}^{(1)} = \prod_{i=1}^{n} \widetilde{p}_{\psi_i}^{(1)}$ (resp. $\widetilde{p}_{\psi}^{(2)} = \prod_{i=1}^{n} \widetilde{p}_{\psi_i}^{(2)}$) defined in eq. (7) (resp. eq. (9)), its expectation $\widetilde{\mathbb{E}}^{(1)}$ (resp. $\widetilde{\mathbb{E}}^{(2)}$) and its ELBO $J^{(1)}(\psi, \theta)$ (resp. $J^{(2)}(\psi, \theta)$) detailed in the next proposition.

**Proposition 3.** *The ELBO defined in eq. (6) with variational approximation $\widetilde{p}_{\psi}^{(1)}$ can be written in matrix form as*

$$\begin{aligned} J^{(1)}(\psi, \theta) =& \widetilde{\mathbb{E}}^{(1)} \left[\log p_\theta(\mathbf{Y}|\mathbf{Z}, \mathbf{W})\right] + \widetilde{\mathbb{E}}^{(1)} \left[\log p_\theta(\mathbf{W})\right] + H(\mathbf{P}) + \frac{np}{2} + \frac{1}{2} \operatorname{Tr}\left(\mathbf{1}_{n,p}^\top \log(\mathbf{S^2})\right) \\ &+ \frac{n}{2} \log|\boldsymbol{\Omega}| - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Omega}\left(\operatorname{Diag}(\bar{\mathbf{S}}^2) + g\left(\mathbf{M} - \mathbf{XB}\right)\right)\right) \end{aligned}$$

*and with variational approximation $\widetilde{p}_{\psi}^{(2)}$ we get*

$$\begin{aligned} J^{(2)}(\psi, \theta) =& \widetilde{\mathbb{E}}^{(2)} \left[\log p_\theta(\mathbf{Y}|\mathbf{Z}, \mathbf{W})\right] + \widetilde{\mathbb{E}}^{(2)} \left[\log p_\theta(\mathbf{W})\right] + H(\mathbf{P}) + \frac{np}{2} + \frac{1}{2} \operatorname{Tr}\left(\mathbf{Q}^\top \log(\mathbf{S^2})\right) \\ &+ \frac{n}{2} \log|\boldsymbol{\Omega}| - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Omega}\left(\operatorname{Diag}\left(\mathbf{1}_n^\top\left(\mathbf{Q} \odot \mathbf{S}^2\right)\right) + g\left(\mathbf{Q} \odot (\mathbf{M} - \mathbf{XB})\right)\right)\right) \\ &- \frac{1}{2} \operatorname{Tr}\left(\operatorname{diag}(\boldsymbol{\Omega}) 1_n^\top \left((1_n \operatorname{diag}(\boldsymbol{\Sigma})^\top) \odot \mathbf{P} + \mathbf{P} \odot \mathbf{Q} \odot (\mathbf{M} - \mathbf{XB})^2\right)\right) \\ &- \frac{1}{2} 1_n^\top \mathbf{P} \log(\operatorname{diag}(\boldsymbol{\Sigma})), \end{aligned}$$

*where $\odot$ denotes the Hadamard product,* diag *returns a vector constituted of the diagonal of the input squared matrix, $\mathbf{1}_n$ is a column-vector of size $n$ filled with 1s, $\mathbf{1}_{n,p}$ is a matrix of size $n \times p$ filled with 1s,* Diag *takes a vector $x$ and returns a diagonal matrix with diagonal $x$, logarithm and squared functions are applied component-wise, $\mathbf{Q} = \mathbf{1}_{n,p} - \mathbf{P}$ and $g(\mathbf{D}) = \mathbf{D}^\top \mathbf{D}$ for $\mathbf{D} \in \mathbb{R}^{n \times p}$. We denoted $\bar{\mathbf{S}}^2 = \mathbf{1}_n^\top \mathbf{S}^2$ and $\delta_{0,\infty}(x) = \begin{cases} 0 & \text{if } x = 0 \\ -\infty & \text{else} \end{cases}$ with the convention that $0 \times \delta_{0,\infty}(x) = 0$ for all $x$ and $0 \times \log(0) = 0$. Note that both ELBOs share the following terms ($\widetilde{\mathbb{E}}^{(1)}$ and $\widetilde{\mathbb{E}}^{(2)}$ coincides for the following terms so that we drop the index):*

$$\begin{aligned} \tilde{\mathbb{E}}\left[\log p_\theta(\mathbf{Y}|\mathbf{Z}, \mathbf{W})\right] &= \operatorname{Tr}\left(\mathbf{Q}^\top \left(\mathbf{Y} \odot (\mathbf{O} + \mathbf{M}) - \mathbf{A} - \log(\mathbf{Y}!)\right) + \mathbf{P}^\top \delta_{0,\infty}\left(\mathbf{Y}\right)\right), \\ \tilde{\mathbb{E}}\left[\log p_\theta(\mathbf{W})\right] &= \operatorname{Tr}\left(\mathbf{P}^\top \mu_0 - \mathbf{1}_{n,p}^\top \log\left(\mathbf{1}_{n,p} + e^{\mu_0}\right)\right), \\ H(\mathbf{P}) &= -\operatorname{Tr}\left(\mathbf{P}^\top \log(\mathbf{P}) + \mathbf{Q}^\top \log\left(\mathbf{Q}\right)\right), \end{aligned}$$

*where factorial and exponential are applied component-wise and the matrix* $\mathbf{A}$ *denotes* $\exp(\mathbf{O} + \mathbf{M} + \mathbf{S}^2/2)$ *where* exp *is applied component-wise and* $\mu_0 = \mathbf{1}_{n,p} \times \text{logit}(\pi)$ *in the ND case,* $\mu_0 = \boldsymbol{X}^0\boldsymbol{B}^0$ *in the CD case and* $\mu_0 = \bar{\boldsymbol{B}}^0\bar{\boldsymbol{X}}^0$ *in the RD case.*

# 4 Optimization

Estimating $\theta$ is equivalent to solving the optimization problem

$$\arg\max_{\psi,\theta} J(\psi, \theta). \tag{10}$$

where $J$ can be either $J^{(1)}$ (standard approximation) or $J^{(2)}$ (enhanced approximation).

## 4.1 Optimization of $J^{(1)}$

Past experience for standard PLN models [Chiquet et al., 2017, 2019, 2021] (and analytical properties of $J^{(1)}$ derived in this section) suggests solving the above problem using alternated gradient descent.

**Proposition 4.** *[Updates of* $\mathbf{B}, \mathbf{\Omega}, \mathbf{P}$ *and* $\mathbf{B}^0$*] For fixed* $\psi$*, the values of* $\mathbf{\Omega}, \mathbf{B}$ *maximizing* $J^{(1)}$ *are*

$$\widehat{\mathbf{\Omega}} = n\left[g(\boldsymbol{M} - \mathbf{XB}) + \bar{\boldsymbol{S}}^2\right]^{-1}, \qquad \widehat{\mathbf{B}} = [\mathbf{X}^\intercal\mathbf{X}]^{-1}\mathbf{X}^\intercal\mathbf{M}.$$

*where* $g(\mathbf{D}) = \mathbf{D}^\intercal\mathbf{D}$ *as in Proposition 3. Furthermore, if* $\mathbf{X}^0 = \mathbf{1}_n$ *,* $J^{(1)}$ *is maximized at* $\widehat{\mathbf{B}}^0 = \frac{1}{n}\mathbf{1}_n^\top\mathbf{P}$*. Likewise, when* $\theta$ *is fixed,* $J^{(1)}$ *is concave with respect to* $\mathbf{P}$ *and maximized at* $\widehat{\mathbf{P}} = \text{logit}^{-1}\left(\mathbf{A} + \mathbf{X}^0\mathbf{B}^0\right) \times \delta_0(\mathbf{Y})$.

Convergence of Algorithm 1 to a stationnary point of $J^{(1)}$ is a direct consequence of the following lemma.

**Lemma 1** (Convergence properties)**.** $J^{(1)}$ *is (separately) concave in* $\theta$ *and* $\psi$*.*

## 4.2 Optimization of $J^{(2)}$

While optimization of $J^{(1)}$ is easily manageable using closed forms and benefits from a bi-concavity property, optimization of $J^{(2)}$ is more challenging. Indeed, the concavity in $\mathbf{\Omega}$ is lost and no closed form can be used for any parameter update.

We do not maximize the ELBO with respect to each parameter in an alternate coordinate-wise fashion but instead compute the gradient with respect to $(\psi, \theta)$ as if it were a single parameter. Formally, given $\psi^{(0)}, \theta^{(0)}$ and a learning rate $\eta > 0$, we perform the update step until an arbitrary criterion is reached or a number of iterations is done.

$$(\psi^{(s+1)}, \theta^{(s+1)}) = (\psi^{(s)}, \theta^{(s)}) + \eta\nabla_{\psi,\theta}J^{(2)}(\psi^{(s)}, \theta^{(s)}) \tag{11}$$

**Algorithm 1:** VEM

**Input** : $\theta^{(0)}, \psi^{(0)}$ initial point, $T \geqslant 1$ number of iterations.

**for** $s = 0, \dots T-1$ **do**

**M-step**

$$\boldsymbol{\Omega}^{(s+1)} = n \left[ g\left( \boldsymbol{M}^{(s)} - \mathbf{X}\mathbf{B}^{(s)} \right) + \bar{\boldsymbol{S}}^{2(s)} \right]^{-1}$$

$$\mathbf{B}^{(s+1)} = [\mathbf{X}^\intercal \mathbf{X}]^{-1} \mathbf{X}^\intercal \mathbf{M}^{(s)}$$

$$\mathbf{B}^{0,(s+1)} = \arg\max_{\mathbf{B}^0} \operatorname{Tr}\left[ \left(\mathbf{P}^{(s)}\right)^\intercal \mathbf{X}^0 \mathbf{B}^0 \right] - \operatorname{Tr}\left[ \mathbf{1}_{n,p}^\intercal \log\left(1 + e^{\mathbf{X}^0 \mathbf{B}^0}\right) \right]$$

**VE-step**

$$\mathbf{P}^{(s+1)} = \operatorname{logit}^{-1}\left( \mathbf{A}^{(s)} + \mathbf{X}^0 \mathbf{B}^{0,(s+1)} \right) \times \delta_0(\mathbf{Y}), \quad \mathbf{Q}^{(s+1)} = \mathbf{1}_{n,p} - \mathbf{P}^{(s+1)}$$

$$\mathbf{M}^{(s+1)} = \arg\max_{\mathbf{M}} \left( \operatorname{Tr}\left( (\mathbf{Y} \odot \mathbf{M} - \mathbf{A})^\top \mathbf{Q}^{(s+1)} \right) - \frac{1}{2} \operatorname{Tr}\left( \boldsymbol{\Omega}^{(s+1)} g\left( \mathbf{M} - \mathbf{X}\mathbf{B}^{(s+1)} \right) \right) \right)$$

$$\mathbf{S}^{(s+1)} = \arg\max_{\mathbf{S}} \left( -\operatorname{Tr}\left( \mathbf{A}^\top \mathbf{Q}^{(s+1)} \right) - \frac{1}{2} \operatorname{Tr}\left( \mathbf{1}_{n,p}^\intercal \log\left( \mathbf{S}^2 \right) \right) - \frac{1}{2} \operatorname{Tr}\left( \boldsymbol{\Omega}^{(s+1)} \bar{\mathbf{S}}^2 \right) \right)$$

**end**

**Output :** $\theta^{(T)}, \psi^{(T)}$

## 4.3 Optimization using analytic law of $W_{ij}|Y_{ij}$

The exact conditional law $W_{ij}|Y_{ij}$ can be derived and is detailed in the next proposition.

**Proposition 5.** *Let $1 \leqslant j \leqslant p$. The conditional law of $W_{ij}|Y_{ij}$ is given by*

$$W_{ij}|Y_{ij} \sim \mathcal{B}\left( \frac{\pi_{ij}}{\varphi\left(\mathbf{X}_i^\top \boldsymbol{B}_j, \Sigma_{jj}\right)(1-\pi_{ij}) + \pi_{ij}} \right) \mathbf{1}_{Y_{ij}=0}$$

*with $\varphi(\mu, \sigma^2) = \mathbb{E}\left[\exp(-X)\right]$, $X \sim \mathcal{LN}(\mu, \sigma^2)$.*

In Section 3 we made the variational approximation $\widetilde{p}(W_{ij}) \sim \mathcal{B}(P_{ij})$, considered $P_{ij}$ as free and optimized the ELBO with respect to $P_{ij}$. The above proposition suggests that $P_{ij}$ can instead be derived directly from $\theta$ and not considered as a free variational parameter. We consider $\widetilde{J}^{(1)}$ (resp. $\widetilde{J}^{(2)}$) the ELBO $J^{(1)}$ (resp. $J^{(2)}$) with $P_{ij} = \Psi(\theta)_{ij}$ with

$$\Psi(\theta) \triangleq \frac{\pi}{\varphi\left(\mathbf{X}^\top \mathbf{B}, \mathbf{1}_n \operatorname{diag}(\boldsymbol{\Sigma})^\top\right) \odot (1-\pi) + \pi} \odot \mathbf{1}_{\boldsymbol{Y}=\mathbf{0}},$$

where $\varphi$ and the division are applied component-wise and $\mathbf{1}_{\boldsymbol{Y}=\mathbf{0}}$ is a $n \times p$ matrix such that $(\mathbf{1}_{\boldsymbol{Y}} = 0)_{ij} = 0$ if and only if $Y_{ij} = 0$. Formally, we have

$$\widetilde{J}^{(1)}(\mathbf{M}, \mathbf{S}, \boldsymbol{\Omega}, \mathbf{B}, \mathbf{B}^0) = J^{(1)}\left(\mathbf{M}, \mathbf{S}, \Psi(\boldsymbol{\Omega}, \mathbf{B}, \mathbf{B}^0), \boldsymbol{\Omega}, \mathbf{B}, \mathbf{B}^0\right),$$

and the same formula applies to $\widetilde{J}^{(2)}$. Note that both $\widetilde{J}^{(1)}$ and $\widetilde{J}^{(2)}$ have $np$ fewer variational parameters compared to $J^{(1)}$ and $J^{(2)}$ ($2np$ compared to $3np$) since $\mathbf{P}$ is now completely determined by $\theta$. The function $\varphi$ is intractable but a sharp (derivable) approximation $\tilde{\varphi}$ is available and detailed in the next section. A major drawback of this approach compared to optimizing $J^{(1)}$ is the lack of any closed form update as stationary points of $\tilde{\varphi}$ are intractable. For the optimization, we consider the gradient scheme defined in Equation (11) where $\psi$ is replaced with $\psi_1 = (\mathbf{M}, \mathbf{S})$.

# 5 Simulation Study

## 5.1 Experimental details

We evaluate model 3a on simulated data. We set $p = 300$ and $d = 1$. Given $\theta^\star = (\mathbf{\Omega}^\star, \mathbf{B}^\star, \pi^\star)$, we simulate $n = 1000$ independent observations $\mathbf{Y}_i$, optimize each criterion, repeat this procedure 10 times and evaluate the Root Mean Squared Error (RMSE) between $\mathbf{\Omega}^\star$ and $\widehat{\mathbf{\Omega}}$, between $\mathbf{B}^\star$ and $\widehat{\mathbf{B}}$ and between $\pi^\star$ and $\widehat{\pi}$. The resuls are displayed in Figure 1.

**Parameters simulation** We employ a simulation approach to generate the matrix $\mathbf{\Sigma}^\star = \mathbf{\Omega}^\star$ in a block-wise structure, where the identity matrix is added on top of it to ensure invertibility. The matrices $\mathbf{X}$ and $\mathbf{B}^\star$ are simulated so that each term of $\mathbf{XB}^\star$ follows a gaussian distribution with unit variance. The mean is chosen along $\{0, 0.5, \ldots, 4\}$ and displayed on the x-axis. The parameter $\pi^\star$ is set to 0.3.

# References


Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, and Géraldine Pascal. FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, 34(8):1287–1294, 12 2017. ISSN 1367-4803.

J. Aitchison and C. H. Ho. The multivariate poisson-log normal distribution. *Biometrika*, 76 (4):643–653, 1989.

Jenni Niku, Francis KC Hui, Sara Taskinen, and David I Warton. gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, 10(12):2173–2182, 2019.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. 39:1–38, 1977.

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.

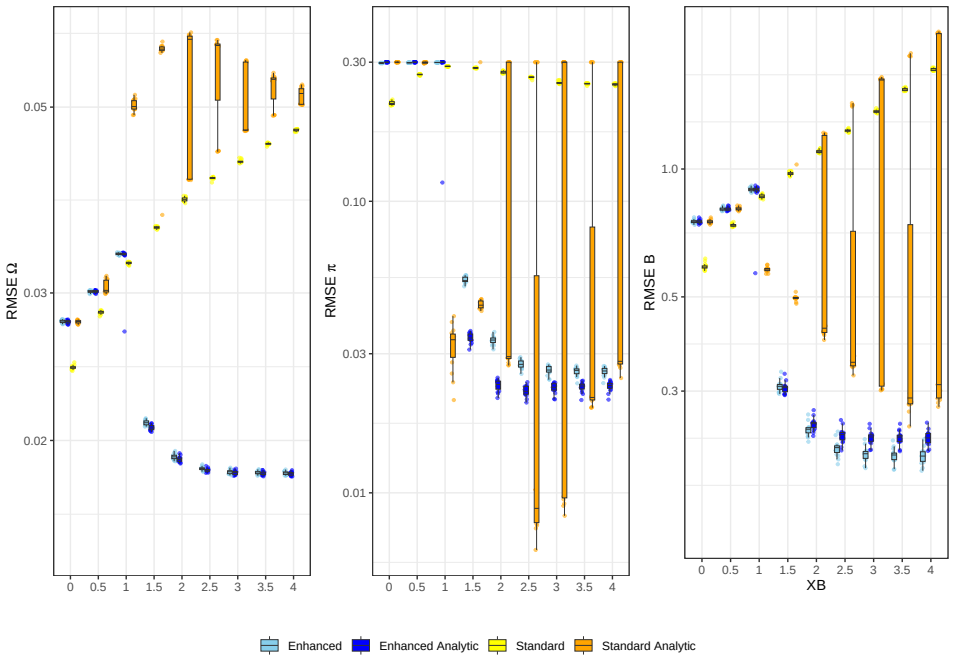



Figure 1: Simulations with $n = 1000, p = 300$ and $d = 1$. Enhanced corresponds to $J^{(2)}$, Enhanced Analytic to $\widetilde{J}^{(2)}$, Standard to $J^{(1)}$ and Standard Analytic to $\widetilde{J}^{(1)}$.

Francis KC Hui, David I Warton, John T Ormerod, Viivi Haapaniemi, and Sara Taskinen. Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26(1):35–43, 2017.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Mcmc maximum likelihood for latent state models. *Journal of Econometrics*, 137(2):615–640, 2007. ISSN 0304-4076.

ISSN 03036898, 14679469.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.

Chin-Shang Li. Identifiability of zero-inflated Poisson models. *Brazilian Journal of Probability and Statistics*, 26(3):306 – 312, 2012.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for probabilistic poisson pca. *Ann. Appl. Stat.*, March 2017.

Julien Chiquet, Stephane Robin, and Mahendra Mariadassou. Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171. PMLR, 2019.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:188, 2021.

# Statistique et sport 1

# Investigating swimming technical skills by a double partition clustering of multivariate functional data allowing for dimension selection

Antoine Bouvet [1,2,3] & Salima El Kolei [3] & Matthieu Marbac [3]

[1] *Univ. Rennes 2, ENS Rennes, M2S Laboratory-EA 7470, France, antoine.bouvet@ens-rennes.fr*
[2] *Inria Rennes Bretagne Atlantique, MIMETIC, France*
[3] *Univ. Rennes, Ensai, CNRS, CREST—UMR 9194, France, salima.el-kolei@ensai.fr, matthieu.marbac-lourdelle@ensai.fr*

**Résumé.** Monitorer les compétences techniques des nageurs constitue un défi majeur en sciences du sport afin d'améliorer les performances. Cela peut être fait en analysant les données fonctionnelles multivariées mesurées par des capteurs miniaturisés tels que les centrales inertielles (IMU). Ces données sont composées de six dimensions décrivant la cinématique 3D du nageur au cours du temps à travers les accélérations et la vitesse angulaire. Pour investiguer les niveaux techniques en crawl sur la base des enregistrements IMU, un modèle de mélange produisant deux partitions complémentaires est proposé et reflète, pour chaque nageur, son pattern de nage et sa capacité à le reproduire. Contrairement aux approches habituelles de clustering de données fonctionnelles, celle-ci prend également en compte les informations présentes dans les termes d'erreur résultant de la décomposition en bases fonctionnelles. En effet, après avoir décomposé en bases fonctionnelles avec un nombre fini d'éléments à la fois le signal original (mesurant le pattern de nage) et le signal des termes d'erreur au carré (mesurant la capacité à le reproduire), la méthode ajuste la distribution conjointe des coefficients liés aux deux décompositions en tenant compte de la dépendance entre les deux partitions. La modélisation de cette dépendance est obligatoire puisque la difficulté à reproduire un pattern de nage dépend de sa forme. En outre, une décomposition éparse de la distribution au sein des composantes permet de sélectionner les dimensions pertinentes lors du clustering. Cela permet d'améliorer l'interprétation technique du modèle pour les utilisateurs. Les partitions obtenues sur les données IMU agrègent la variabilité cinématique de la nage associée aux compétences techniques et permettent d'identifier les habiletés biomécaniques pertinentes pour des sprinteurs en crawl.

**Mots-clés.** Clustering, Sélection de variables, Données fonctionnelles, Performance sportive

**Abstract.** Investigating technical skills of swimmers is a challenge for sports science to reach performance improvement. It can be achieved by analyzing multivariate functional data recorded by miniaturized sensors such as Inertial Measurement Units (IMU). These data are composed of six dimensions describing swimmer's kinematic through the 3D accelerations and angular velocity temporal records. To investigate technical levels of front-crawl swimmers, a new model-based approach is introduced to obtain two complementary partitions reflecting, for each swimmer, its swimming pattern and its ability to reproduce it based on the IMU records. Contrarily to the usual approaches for functional data clustering, the proposed approach also considers the information of the error terms resulting from the functional

basis decomposition. Indeed, after decomposing into functional basis with finite number of elements both the original signal (measuring the swimming pattern) and the signal of squared error terms (measuring the ability to reproduce it), the method fits the joint distribution of the coefficients related to both decompositions by considering dependency between both partitions. Modeling this dependency is mandatory since the difficulty of reproducing a swimming pattern depends on its shape. Moreover, a sparse decomposition of the distribution within components that permits a selection of the relevant dimensions during clustering is proposed. It allows the improvement of the technical interpretation of the model for users. The partitions obtained on the IMU data aggregate the kinematical stroke variability linked to swimming technical skills and allow relevant biomechanical ability for front-crawl sprinter to be identified.

**Keywords.** Clustering, Feature selection, Functional data, Sport performance

# 1 Introduction

Tracking of technical skills by comprehensive training monitoring is a main challenge for sport performance improvement, especially in swimming that requires efficient movement in aquatic environment. Swimming technique is defined by the repetition of similar but not identical stroke patterns constituted of instabilities called biomechanical variability (Fernandes et al. (2022a)). Stroke variability plays a major role in generating swimming speed because it is related to swimming efficiency and differs according to swimming performance levels (Seifert et al. (2016)). Thus, technical skills can be described by this biomechanical variability and quantified through two complementary and associated aspects of motion: the swimming pattern and the ability to reproduce it (Fernandes et al., 2022b). Development of automatic methodologies supporting on-board investigation of both of these components is promising. To do this, Inertial Measurement Units (IMU) are used to provide embedded kinematical data collection regarding sports related movements through tri-axial accelerometer and gyroscopic temporal records. However, the current literature is focused on empirical indicators describing stroke patterns without taking their functional nature into account. Hence, this leads to limited insights regarding underlying kinematical variability defining technical skills and making their conclusions weak or not representative. Their main limitation is due to the absence of a powerful statistical model able to analyze the complex multivariate dynamics of swimming patterns.

Statistical analysis of multivariate functional periodic IMU data supporting monitoring of kinematical variability could be achieved by a dependent double partition clustering measuring the swimming pattern (*i.e.,* first latent partition) and the ability to reproduce it (*i.e.,* second latent partition). Indeed, the type of stroke pattern directly impacts its repeatability. Moreover, dimension selection during this double clustering could allow the IMU axes to be identified for explaining disparities between swimmers and associated technical skills.

Traditionally, statistical methods allowing functional data to be clustered consider a decomposition of the data into a functional basis, and then use classical multivariate methods for

finite dimensional data directly on these basis coefficients. Considering this approach, different model-based clustering methods have been developed for univariate functional data as well as for multivariate functional data (Bouveyron et al. (2019). In sports science, model-based clustering approaches for functional data provide a useful framework for new insights on performance (Leroy (2020)) that can outpace classical approaches relying on empirical analysis that lead to limited kinematical information (Mallor et al., 2010). For all the methods considering functional basis decompositions, the loss of information due to the approximation of the original data into a functional basis is neglected. However, for investigating swimming techniques, it is crucial to keep the information of the dispersion around the swimming pattern that is traditionally lost by using the basis decomposition.
Some model-based clustering approaches have been developed to perform a selection of the variables (Marbac and Sedki, 2017). Selecting variables is very challenging in clustering because the role of a variable is defined with respect to a variable that is not observed. Thus, the selection of the variables and the clustering need to be performed simultaneously. In the context of clustering multivariate functional data, to the best of our knowledge, no methods can lead to a detection of the dimensions that are relevant for clustering. However, one can consider a natural extension of the model-based clustering approach performing feature selection. Indeed, when the functional data are decomposed into a functional basis, claiming that a dimension of the functional data is not relevant for clustering means that all the coefficients of the functional basis, related to this dimension, are not informative for the clustering.
We propose a double partition clustering approach to investigate swimming technical skills using IMU data that are periodic due to the repetition of strokes. Moreover, the proposed model-based approach allows for the identification of the discriminative dimensions for each partition. After decomposing both of the original signals (*i.e.,* measuring the swimming pattern) and the signals of squared error terms (*i.e.,* measuring the ability to reproduce the swimming pattern) into a Fourier basis, the method fits the joint distribution of the coefficients related to both decompositions by considering dependency between both partitions. Modelling this dependency is important because the difficulty of reproducing a swimming pattern depends on its shape and then on technical skills. The model considers that the information about the swimming pattern that measures the kinematical smoothness is contained in the coefficients arising from the decomposition of the original signal while the information about the ability of reproducing the pattern is contained in the coefficients arising from the decomposition of the signal of the squared error terms. As usual for a standard model-based approach to cluster functional data, a sparse decomposition of the distribution within components is used. Here, we consider a conditional independence between the coefficients of different dimensions within the component. This assumption allows an automatic selection to be made of the relevant dimensions during clustering that improves the accuracy of the estimates and that highlights the dimensions that are discriminative for both partitions.
This paper is organized as follows. Section 2 presents the double clustering model-based framework and dimension selection for multivariate functional data. Identifiability issues of this new model are investigated and presented. Mathematical details, model inference and numerical experiments are available in Bouvet et al. (2024) . Section 3 is devoted to the analysis of the SWIMU data and the consequent biomechanical interpretation for technical swimming skills analysis. Section 4 gives a conclusion.

# 2 Model-based approach for estimating a double partition from multivariate functional data

## 2.1 Mixture model on basis coefficients

We consider a random sample $\mathbf{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$ composed of $n$ independent and identically distributed multivariate time series. Each individual $i$ is described by a $J$-dimensional discrete-time time series $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^\top, \ldots, \boldsymbol{X}_{iJ}^\top)^\top$, where $\boldsymbol{X}_{ij} = (X_{ij}(1), \ldots, X_{ij}(T_i))^\top$ and $X_{ij}(t) \in \mathbb{R}$ denotes the value of dimension $j$ of the multivariate time series measured on subject $i$ at time $t$, $T_i$ being the length of the multivariate time series recorded on subject $i$. Each univariate time series admits a basis expansion leading to

$$X_{ij}(t) = \boldsymbol{Y}_{ij}^\top \boldsymbol{\psi}_j(t) + \varepsilon_{ij}(t), \tag{1}$$

where $\boldsymbol{Y}_{ij} \in \mathbb{R}^{G_j}$ is a $G_j$-variate random variable that groups the $G_j$ basis coefficients, $\boldsymbol{\psi}_j(t) = (\psi_{j1}(t), \ldots, \psi_{jG_j}(t))^\top$ is the vector containing the values of the $G_j$ basis functions evaluated at time $t$ and where the term of random error $\varepsilon_{ij}(t)$ is supposed to be centered given the natural filtration $\mathcal{F}_i(t)$ (*i.e.*, $\mathbb{E}[\varepsilon_{ij}(t) \mid \mathcal{F}_i(t-1)] = 0$). Let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}^\top, \ldots, \boldsymbol{Y}_{iJ}^\top)^\top \in \mathbb{R}^G$ be the vector of length $G = \sum_{j=1}^J G_j$ that gathers the basis coefficients of subject $i$ for the $J$ dimensions and let $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}^\top, \ldots, \boldsymbol{\varepsilon}_{iJ}^\top)^\top$ be the vector of the error terms of individual $i$ where $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij}(1), \ldots, \varepsilon_{ij}(T_i))^\top$. In addition, we express each univariate time series of the squared error term in a functional basis with $H_j$ elements as follows

$$\varepsilon_{ij}^2(t) = \boldsymbol{Z}_{ij}^\top \boldsymbol{\Pi}_j(t) + \xi_{ij}(t), \tag{2}$$

where $\boldsymbol{Z}_{ij} \in \mathbb{R}^{H_j}$ is a $H_j$-variate random variable that groups the $H_j$ basis coefficients, $\boldsymbol{\Pi}_j(t) = (\Pi_{j1}(t), \ldots, \Pi_{jH_j}(t))^\top$ is the vector containing the values of the $H_j$ basis functions evaluated at time $t$, $\mathbb{E}[\xi_{ij}(t) \mid \mathcal{F}_i(t-1)] = 0$.
The decomposition defined by (1)-(2) considers two general functional basis with intrinsic dimension assumed to be finite. To model the cyclical pattern of the swimmers' motion, it seems appropriate to use a Fourier basis for the decomposition of each time series. The period is the same among the dimensions of the functional data but the degrees of the Fourier basis can be different. Also, although the period of the basis differs for each swimmer due to the different swimming periods between the swimmers. This follows that, if swimmer $i$ has the same period $\zeta_i$, we have $\boldsymbol{\psi}_j(t) := \boldsymbol{\psi}_j(t;\zeta_i)$ where $\boldsymbol{\psi}_j(t;\zeta_i) = (\boldsymbol{\psi}_{j1}(t;\zeta_i), \ldots, \boldsymbol{\psi}_{jG_j}(t;\zeta_i))^\top$, $\boldsymbol{\psi}_{j1}(t;\zeta_i) = 1$ and $\boldsymbol{\psi}_{j(2\ell)}(t;\zeta_i) = \cos(2\pi\ell t/\zeta_i)$ and $\boldsymbol{\psi}_{j(2\ell+1)}(t;\zeta_i) = \sin(2\pi\ell t/\zeta_i)$, for $\ell = 1, \ldots, (G_j-1)/2$, and we have $\boldsymbol{\Pi}_j(t) := \boldsymbol{\Pi}_j(t;\zeta_i)$ where $\boldsymbol{\Pi}_j(t;\zeta_i) = (\boldsymbol{\Pi}_{j1}(t;\zeta_i), \ldots, \boldsymbol{\Pi}_{jH_j}(t;\zeta_i))^\top$, $\boldsymbol{\Pi}_{j1}(t;\zeta_i) = 1$ and $\boldsymbol{\Pi}_{j(2\ell)}(t;\zeta_i) = \cos(2\pi\ell t/\zeta_i)$ and $\boldsymbol{\Pi}_{j(2\ell+1)}(t;\zeta_i) = \sin(2\pi\ell t/\zeta_i)$, for $\ell = 1, \ldots, (H_j-1)/2$.
We aim at grouping the swimmers according to their swimming patterns as well as their abilities to reproduce them. This implies the estimation of two latent categorical variables $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{iK})^\top$ and $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{1L})^\top$ that indicate the swimming pattern and the dispersion around this pattern for swimmer $i$ respectively, $K$ and $L$ denoting the number of different swimming patterns and the number of types of dispersion around a pattern. Since the basis coefficients $\boldsymbol{Y}_i$ depend on the mean behavior of $\boldsymbol{X}_i$, we consider that this vector contains all the information about $\boldsymbol{V}_i$. Similarly, since the basis coefficients $\boldsymbol{Z}_i$ depend on the

squared of the error terms $\boldsymbol{\varepsilon}_i$, we consider that this vector contains all the information about $\boldsymbol{W}_i$. Thus, the model assumes conditional independence between $\boldsymbol{Y}_i$ and $\boldsymbol{W}_i$ given $\boldsymbol{V}_i$ and conditional independence between $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{V}_i$ given $\boldsymbol{W}_i$. Finally, it allows for dependency between $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ and it assumes conditional independence between $\boldsymbol{Y}_i$ and $\boldsymbol{\varepsilon}_i$ given $\boldsymbol{W}_i$ and $\boldsymbol{V}_i$. Thus, the basis coefficients follow a specific mixture model with $K \times L$ components defined by the following probability distribution function (pdf)

$$p(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \pi_{k\ell} f_k(\boldsymbol{y}_i; \boldsymbol{\alpha}_k) g_\ell(\boldsymbol{z}_i; \boldsymbol{\beta}_\ell), \tag{3}$$

where $\pi_{\ell k} := \mathbb{P}(V_{ik} = 1, W_{i\ell} = 1) > 0$, $\sum_{k=1}^{K} \sum_{\ell=1}^{L} \pi_{k\ell} = 1$, $\boldsymbol{\theta}$ groups all parameters of model $\boldsymbol{m}$, $f_k$ is the pdf of cluster $k$ for the swimming pattern parameterized by $\boldsymbol{\alpha}_k$ that defines the conditional distribution of $\boldsymbol{Y}_i$ given $V_{ik} = 1$ and $g_\ell$ is the pdf of cluster $\ell$ for the ability of reproducing the swimming pattern parameterized by $\boldsymbol{\beta}_\ell$ that defines the conditional distribution of $\boldsymbol{Z}_i$ given $W_{i\ell} = 1$.
The model defined by (3) is a parsimonious mixture model that imposes equality constraints between the parameters of some pdfs of its components. It implies that the marginal distribution of $\boldsymbol{Y}_i$ is a mixture model with $K$ components and that the marginal distribution of $\boldsymbol{Z}_i$ is a mixture model with $L$ components. Note that dependency between $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ is considered by (3) and thus this model is not equivalent to a product of two mixture models modeling the marginal distributions of $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$. Finally, we present two properties of the model. Detailed proofs for lemmas 1 and 2 are presented in Bouvet et al. (2024).

**Lemma 1.** *If the parameters of the marginal distributions $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ are identifiable, then the parameters of model defined by* (3) *are identifiable.*

The second property states that considering the dependency between the two latent partitions, and thus using the joint distribution of $(\boldsymbol{Y}_i^\top, \boldsymbol{Z}_i^\top)^\top$ for estimating the two partitions leads to a better estimator of both partitions than considering an estimator of $\boldsymbol{V}_i$ based on $\boldsymbol{Y}_i$ and an estimator of $\boldsymbol{W}_i$ based on $\boldsymbol{Z}_i$ when $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ are not independent. Let $\Upsilon_V(\boldsymbol{Y}_i)$ and $\Upsilon_V(\boldsymbol{Y}_i, \boldsymbol{Z}_i)$ be the applications that associate an estimator of $\boldsymbol{V}_i$ (*i.e.,* a vector of length $K$ composed by zeros except for one coordinate that is equal to one) by using the *MAP* rule (*i.e.,* affecting an observation to the most likely cluster) based on the distribution of $\boldsymbol{Y}_i$ and $(\boldsymbol{Y}_i^\top, \boldsymbol{Z}_i^\top)^\top$ respectively. Let $\Upsilon_W(\boldsymbol{Z}_i)$ and $\Upsilon_W(\boldsymbol{Y}_i, \boldsymbol{Z}_i)$ be the applications that associate an estimator of $\boldsymbol{W}_i$ (*i.e.,* a vector of length $L$ composed by zeros except for one coordinate that is equal to one) by using the *MAP* rule based on the distribution of $\boldsymbol{Z}_i$ and $(\boldsymbol{Y}_i^\top, \boldsymbol{Z}_i^\top)^\top$ respectively.

**Lemma 2.** *If the model* (3) *holds true and if the $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ are not independent then, under the assumption of Lemma 1 and if the densities $f_k$ and $g_\ell$ are continuous for any $k$ and $\ell$, then*

$$\mathbb{E}[\Upsilon_V(\boldsymbol{Y}_i, \boldsymbol{Z}_i) \neq \boldsymbol{V}_i] < \mathbb{E}[\Upsilon_V(\boldsymbol{Y}_i) \neq \boldsymbol{V}_i] \text{ and } \mathbb{E}[\Upsilon_W(\boldsymbol{Y}_i, \boldsymbol{Z}_i) \neq \boldsymbol{W}_i] < \mathbb{E}[\Upsilon_W(\boldsymbol{Z}_i) \neq \boldsymbol{W}_i].$$

## 2.2 Parsimonious model for detecting the relevant dimensions

Since the decompositions into the functional basis can produce high-dimensional vectors, it is usual to assume parsimonious constraints on the dependency within components. Here, we consider that the mixture components belong to the same parametric family and that the coefficients related to different dimensions are conditionally independent given the latent variables. This leads to $\boldsymbol{Y}_{ij} \perp \boldsymbol{Y}_{ij'}|\boldsymbol{V}_i$ and $\boldsymbol{Z}_{ij} \perp \boldsymbol{Z}_{ij'}|\boldsymbol{W}_i$, for $j \neq j'$. We denote by $\phi_j(\cdot;\boldsymbol{\alpha}_{kj})$ the $G_j$-dimensional density of $\boldsymbol{Y}_{ij}$ within component $k$ parameterized by $\boldsymbol{\alpha}_{kj}$ and by $\varphi_j(\cdot;\boldsymbol{\beta}_{\ell j})$ the $H_j$-dimensional density of $\boldsymbol{Z}_{ij}$ within component $\ell$ parameterized by $\boldsymbol{\beta}_{\ell j}$. Therefore, we have

$$f_k(\boldsymbol{y}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{J} \phi_j(\boldsymbol{y}_{ij};\boldsymbol{\alpha}_{kj}) \text{ and } g_\ell(\boldsymbol{z}_i;\boldsymbol{\beta}_\ell) = \prod_{j=1}^{J} \varphi_j(\boldsymbol{z}_{ij};\boldsymbol{\beta}_{\ell j}). \tag{4}$$

The model can consider any parametric multivariate density for $\phi_j$ and $\varphi_j$ and thus allows the usual parametric assumptions to be made to cluster functional data based on the coefficients of their basis extension. The main benefits of (4) is that it easily permits a selection of the dimensions that are relevant for clustering in (3) and so allows to extend variable selection methods for clustering to functional data. In our context, a dimension is relevant for estimating one partition if the coefficients related to this dimension do not have the same distribution among the mixture components. Therefore, dimension $j$ is irrelevant for estimating the swimming pattern if $\boldsymbol{\alpha}_{1j} = \ldots = \boldsymbol{\alpha}_{Kj}$ while this dimension is irrelevant for estimating the ability of reproducing the swimming pattern if $\boldsymbol{\beta}_{1j} = \ldots = \boldsymbol{\beta}_{Lj}$. We denote by $\boldsymbol{\Omega} \subseteq \{1,\ldots,J\}$ and $\boldsymbol{\Gamma} \subseteq \{1,\ldots,J\}$ the indexes of the dimensions that are relevant for estimating the swimming pattern and for estimating the ability of reproducing the swimming pattern respectively. Thus, for a fixed model $\boldsymbol{m} = \{K, L, \boldsymbol{\Omega}, \boldsymbol{\Gamma}\}$, using (3)-(4) and the definition of the relevant dimensions, the pdf of the observed data is defined by

$$f(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \left[\prod_{j\in\boldsymbol{\Omega}^c} \phi_j(\boldsymbol{y}_{ij};\boldsymbol{\alpha}_{1j}) \prod_{j'\in\boldsymbol{\Gamma}^c} \varphi_{j'}(\boldsymbol{z}_{ij'};\boldsymbol{\beta}_{1j'})\right] \times \sum_{k=1}^{K}\sum_{\ell=1}^{L} \pi_{k\ell} \prod_{j\in\boldsymbol{\Omega}} \phi_j(\boldsymbol{y}_{ij};\boldsymbol{\alpha}_{kj}) \prod_{j'\in\boldsymbol{\Gamma}} \varphi_{j'}(\boldsymbol{z}_{ij'};\boldsymbol{\beta}_{\ell j'}). \tag{5}$$

The following lemma gives sufficient conditions to state the identifiability of the parameters of the proposed model (5). Detailed proof for lemma 3 is presented in Bouvet et al. (2024).

**Lemma 3.** *If card*$(\boldsymbol{\Omega}) \geq 1$, *card*$(\boldsymbol{\Gamma}) \geq 1$, *exist* $j \in \boldsymbol{\Omega}$ *and* $j' \in \boldsymbol{\Gamma}$ *such that the marginal distribution of* $\boldsymbol{Y}_{ij}$ *and* $\boldsymbol{Z}_{ij'}$ *is identifiable, then the parameters of model* (5) *are identifiable.*

# 3 Analysis of SWIMU data

The SWIMU database used in this study includes $n = 68$ all-out 25m front-crawl from recreational to world-class swimmers. The participants were instrumented with one IMU located

on the sacrum. To deal with the issue of curve alignment, data pre-processing is conducted in order to set the first frame at the beginning of a stroke by zero-crossing on second-order Butterworth band-pass filtered between 0.1 and 1 Hz on mediolateral acceleration. Since the IMU swimming records are periodic, we decompose these multivariate functional data into a Fourier basis. The period is identified using the previously described zero-crossing. We select the degree of the Fourier basis $G_j$ and $H_j$ used for the decompositions (1) and (2) of each dimension $j$ that minimizes the least square error obtained by leave-one-out cross-validation. Thus, the selected degrees are between 12 and 18 for the original signal decomposition and between 8 and 12 for the decomposition of the squared error terms.

In this application, we consider a Gaussian distribution within components with a diagonal matrix *i.e.,* $\phi_j$ is the pdf of Gaussian distribution with mean $\boldsymbol{\mu}_{jk} = (\mu_{jk1}, \ldots, \mu_{jkG_j})^\top$ and a diagonal covariance matrix. Similarly, $\varphi_j$ is the pdf of Gaussian distribution with mean $\boldsymbol{\nu}_{j\ell} = (\nu_{j\ell 1}, \ldots, \nu_{j\ell H_j})^\top$ and a diagonal covariance matrix. Best model according to the BIC is composed of 2 clusters for the swimming pattern (*e.g.,* $K = 2$), 3 clusters for the ability of reproducing the pattern (*e.g.,* $L = 3$). Moreover, five among the six dimensions are selected for the swimming pattern (all the dimensions but the mediolateral angular velocity) and all the dimensions are selected for the ability of reproducing the pattern. Mediolateral angular velocity mainly reflects the pitch movement of the swimmers and so is not a discriminant constitutive motion of technical abilities for front-crawl sprint.

Figure 1 allows for an easy interpretation of the results. Indeed, it summarizes the swimming patterns by presenting in black plain lines, for each dimension $j$, the mean curves defined, for any $k \in \{1, 2\}$ and $t \in [0, 1[$, by $\overline{X}_{jk}(t) := \widehat{\boldsymbol{\mu}}_{jk}^\top \boldsymbol{\psi}_j(t; 1)$. Moreover, the three clusters of abilities for reproducing the swimming pattern are summarized, for each dimension $j$, by the repeatability region, at each time $t$, $[\overline{X}_{jk}(t) - 2\bar{\varepsilon}_{j\ell}(t), \overline{X}_{jk}(t) + 2\bar{\varepsilon}_{j\ell}(t)]$ defined as the area that differs from two standard deviations at each time $t$ from the mean curve where, for any $\ell \in \{1, 2, 3\}$ and $t \in [0, 1[$, $\bar{\varepsilon}^2_{j\ell}(t) := \widehat{\boldsymbol{\nu}}_{j\ell}^\top \boldsymbol{\Pi}_j(t; 1)$.

The partition of swimming patterns is composed of a majority class (81%) presenting a smooth acceleration pattern with continuity of propulsive actions during the stroke cycle. This class, called the *smoothy* class is characterized by a lower acceleration variation on the propulsive axis (*i.e.,* longitudinal and anteroposterior acceleration) and a higher angular velocity variation on both longitudinal and anteroposterior axes in addition to mediolateral acceleration. The second class of swimming patterns (19%) is composed of more explosive and irregular stroke patterns. It includes higher acceleration variations than in the *smoothy* class particularly on the propulsive axis, associated with reduced mediolateral acceleration and angular velocity variations inducing more stability and a better alignment of the body. This steadiness especially occurs during the peak propelling phases of the cycle (*i.e.,* peaks on longitudinal acceleration). This class is then called the *jerky* class.

The partition of abilities to reproduce the swimming patterns is composed of two classes having equal proportions (37%) characterized by moderate and high stroke pattern repeatability, and of a third class (26%) with lower level. We respectively call them the *moderate*, *high* and *low repeatability* classes (see colors in Figure 1). The repeatability region around mean curves are not constant indicating unequal variance. In this way, this functional visualisation provides a useful understanding of kinematical variability during specific stroke cycle phases. The assumption of independence between both partitions is rejected according to a Pearson's Chi-squared test ($\chi^2$=15.62, df=2, $p \leq$0.01). This confirms the biomechanical association

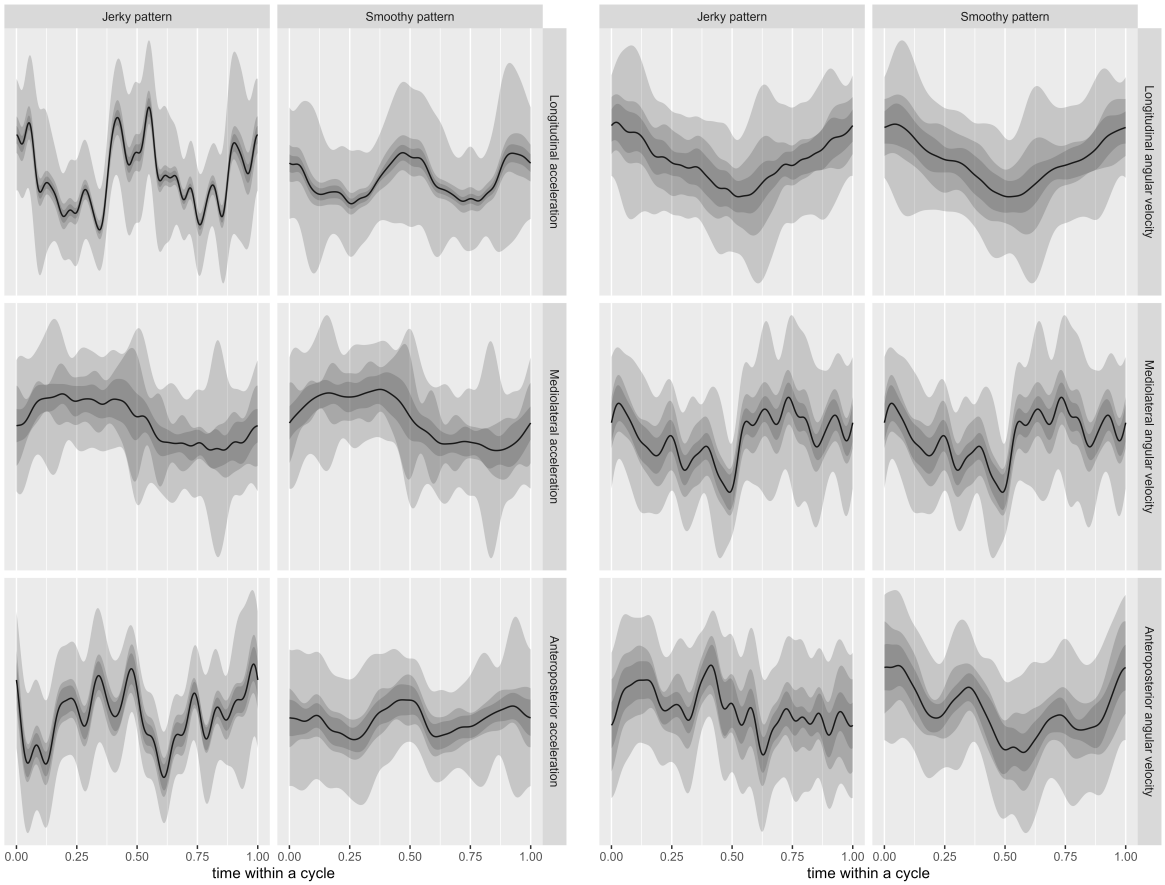



Figure 1: Description of the double partition clustering obtained on the SWIMU data for the six dimensions : columns correspond to the partition defined by the two swimming patterns (*jerky* and *smoothy*) that are represented by their mean curves $\overline{X}_{jk}(t)$ plotted in black plain lines, colors correspond to the partition defined by the three clusters of abilities to reproduce the swimming patterns (dark gray: high repeatability, gray: moderated repeatability, light gray: low repeatability) that are represented by the repeatability regions $[\overline{X}_{jk}(t) - 2\overline{\varepsilon}_{j\ell}(t), \overline{X}_{jk}(t) + 2\overline{\varepsilon}_{j\ell}(t)]$ defined by the dashed lines.

between the swimming pattern and the ability to reproduce it as a discriminant feature of technical skills for front-crawl sprint swimming.
The two partitions clustering allows technical skills to be measured. We now investigate the relation between the technical skills reflected by estimated partitions and performance (see Figure 2). There is a significant effect of gathered partitions on swimming speed as shown by a Fisher test ($F(5.62)=10.1$, $p \leq 0.001$, $\eta^2=0.45$). The *jerky* swimming pattern associated with *low repeatability* is the fastest biomechanical strategy ($1.86 \pm 0.10$ m/s). Indeed, all its pairwise comparisons with others clusters are significant considering a nominal level of 0.05. There are no other significant pairwise comparisons between all kinds of other clusters.
Our approach allows us to discriminate the performance level since there is a clear speed trend for the *jerky+low repeatability* cluster. Thus, it enables to group swimmers of homogeneous technical skills regarding performance. Furthermore this approach allows to perform

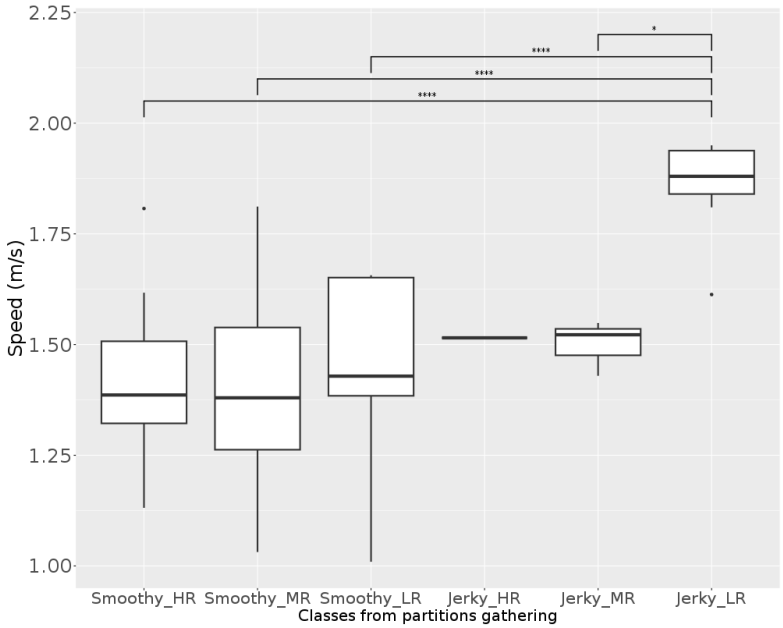



Figure 2: Boxplot of sport performance (*i.e.,* swimming speed) according to gathered partitions from clustering defining swimming pattern and ability to reproduce it. Stars indicate significant pairwise comparisons between classes: * $p \leq 0.05$, **** $p \leq 0.0001$

technical skills evaluation according to stroke cycle kinematics functional modelling on the micro-scale, instead of the classic macro-scale spatio-temporal parameters of literature. In this way, this double partition clustering provides a valuable sports-related outcome.

# 4 Conclusion

We have developed a model-based approach that provides two complementary partitions allowing technical skills to be tracked in swimming based on multivariate functional data. The model considers both the kinematical information from the original signals and the squared error term resulting from the functional basis decomposition and allows for dependency of the two partitions. Otherwise, the method allows for dimension selection to better establish the biomechanical contribution to technical skills.
The results of the application confirm the double partition model's sensitivity to aggregate kinematical variability defining swimming technical skills. Clustering IMU swimming data highlights specific kinds of stroke kinematics related to speed. In this way it allows us to identify a relevant biomechanical ability for front-crawl sprint performance relying on a jerky unstable pattern. Technical skills are driven by management of kinematical variability through, on the one hand, specific swimming patterns linked to stroke smoothness defining continuity of propulsive actions and energy expenditure, and on the other hand, repeatability defining pattern stability and ability to reproduce swimming strokes.
The development of this procedure expands traditional technique monitoring by avoiding only relying on human-based observations of experts or on macro-scale spatio-temporal parame-

ters recorded by a stopwatch. Hence it gives to coaches a complementary data-driven method less subjective to current eyes-based method. Firstly, to better establish skills evaluation in environmental conditions of daily training and secondly to better characterize technical levels of front-crawl swimmers through an automatic user-friendly framework. The proposed model could be applied to a wide population with different characteristics and leads to biomechanical profiling of swimmers. It provides the first modelling of swimming IMU data in the literature.

# References


Bouvet, A., Kolei, S. E., and Marbac, M. (2024). Investigating swimming technical skills by a double partition clustering of multivariate functional data allowing for dimension selection. *Ann. Appl. Statist.*, 18(2):1750–1772.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Fernandes, A., Goethel, M., Marinho, D. A., Mezêncio, B., Vilas-Boas, J. P., and Fernandes, R. J. (2022a). Velocity variability and performance in backstroke in elite and good-level swimmers. *International Journal of Environmental Research and Public Health*, 19(11):6744.

Fernandes, A., Mezêncio, B., Soares, S., Duarte Carvalho, D., Silva, A., Vilas-Boas, J. P., and Fernandes, R. J. (2022b). Intra-and inter-cycle velocity variations in sprint front crawl swimming. *Sports Biomechanics*, pages 1–14.

Leroy, A. (2020). *Multi-task learning models for functional data and application to the prediction of sports performances*. PhD thesis, Université de Paris.

Mallor, F., Leon, T., Gaston, M., and Izquierdo, M. (2010). Changes in power curve shapes as an indicator of fatigue during dynamic contractions. *Journal of biomechanics*, 43(8):1627–1631.

Marbac, M. and Sedki, M. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, 27(4):1049–1063.

Seifert, L., De Jesus, K., Komar, J., Ribeiro, J., Abraldes, J., Figueiredo, P., Vilas-Boas, J., and Fernandes, R. (2016). Behavioural variability and motor performance: Effect of practice specialization in front crawl swimming. *Human movement science*, 47:141–150.

# Analyse d'une compétition mondiale de football féminin par Process Mining.

Laly Lacroix [1] & Julie Treilhou [1] & Emmanuelle Claeys [2] & Sébastien Déjean [3]

[1] *INSA Toulouse, France*
[2] *Unviversité Paul Sabatier, IRIT, Toulouse, France*
[3] *Institut de Mathématiques de Toulouse, Toulouse, France*

**Résumé.** Cet article présente un retour d'expérimentation d'analyse statistique sur des données issues de la *FIFA Women's World Cup 2023*, coupe du monde de football. Contrairement aux indicateurs traditionnels développés en statistique sportive, notre approche intègre l'utilisation d'outils de *process mining*, une méthodologie rarement explorée dans le domaine du sport jusqu'à présent. Cette démarche nous permet d'analyser les processus sous-jacents de manière approfondie, offrant ainsi une perspective nouvelle pour analyser la performance d'une équipe. Les données présentées ici proviennent de la Coupe du monde féminine 2023 représentant des équipes de haut niveau et donnant une visibilité au football féminin. L'intégralité des données sont disponibles sur le site StatsBomb[1], l'une des principales source dans l'*open data* sportif. Les résultats obtenus ont permis une cartographie des trajectoires des tirs ainsi qu'une mise en évidence des schémas tactiques, grâce aux méthodes de *process mining*. Ils offrent également la possibilité d'observer l'évolution tactique de l'équipe victorieuse, en l'occurrence l'Espagne, depuis son premier match jusqu'au dernier.

**Mots-clés.** statistique sportive, event log, sport féminin.

**Abstract.** This article presents the results of a statistical analysis experiment carried out on the Women's World Cup 2023 competition. Unlike traditional indicators developed in sports statistics, our approach are based on *process mining* tools. This approach enables us to analyze underlying processes in plays, offering a new point of view on the analysis of sports performance. The data presented here representing both top-level teams and giving visibility to women's soccer. All the data is available on the StatsBomb website, one of the biggest open-source data platforms. The results provide map shot trajectories and highlight tactical patterns, using *process mining* methods. It also highlights the tactical evolution of the winning team, in this case Spain, from its first to its last match.

**Keywords.** Sports Statistics, event log, women's sport

[1] https://statsbomb.com

# 1 Introduction

L'analyse statistique appliquée au football représente une discipline évolutive et cruciale dans la compréhension approfondie des performances sportives [3]. L'intérêt grandissant pour l'analyse de données dans de nombreuses disciplines permet d'adapter au mieux l'entraînement [4]. Dans le cadre du football, les indicateurs doivent permettre notamment d'extraire les dynamiques tactiques des équipes. Les données sportives présentent des caractéristiques particulièrement intéressantes :

- D'une part les observations utilisées sont parfois issues d'un seul match, le contexte (joueurs présents ce jour là, équipe et période considérée) étant fortement différent d'un match a un autre
- D'autre part les observations peuvent se présenter sous la forme d'un *event log*, c'est-à-dire un jeu de données enregistrant un ensemble d'événements temporels.

Ces propriétés peuvent être en contradiction avec les outils nécessitant des hypothèses telles que des variables indépendantes et identiquement distribuées (i.i.d.) ou un volume substantiel de données. Parmi les métriques fréquemment étudiées, on retrouve la possession de balle, les tirs au but, la précision des passes, ainsi que les statistiques individuelles des joueuses, telles que les dribbles réussis ou encore les interceptions. Bien que ces indicateurs permettent d'évaluer la performance globale d'une équipe, pour enrichir davantage cette analyse, nous proposons d'utiliser une approche novatrice telle que le *process mining* pour obtenir une compréhension plus fine des schémas de jeu, de transitions entre phases de jeu, et des variations tactiques au sein d'une équipe tout au long de la compétition. La première section présente la modélisation de l'*event log* par rapport aux données utilisées, la seconde partie présente les résultats obtenus suite à la comparaison entre deux matchs (ouverture : Costa - Rica vs Espagne et finale : Angleterre vs Espagne). La dernière section conclut sur nos résultats. Nous utilisons principalement la librairie BupaR [2] (codée en `R`), l'intégralité du code pour reproduire les expériences est disponible sur un dépôt GitHub[2].

# 2 Modélisation

L'exploration de processus implique des méthodes d'analyse de *process* représentés par des modèles à partir de *event logs* (c'est-à-dire les données réelles émises, supposées suivre le processus observé) [1]. Chaque événement de l'*event logs* est composé de trois informations : un identifiant de cas, un horodatage et une activité. Grâce à ces informations, un événement (cas) rapporte plusieurs actions réalisées (activités) différenciées par un marqueur temporel (horodatage). Les activités sont rattachées à un même évènement pour

[2]https://github.com/julietrlh/StatbombR

former une séquence, ainsi une séquence individuelle est appelée *trace*. Toutes les traces possibles sont représentées sous forme de DFG (*Direct Follower Graph*). Ce graphique décrivant tous les parcours possibles est appelé *process map*.

Un exemple d'*event log* traditionnellement utilisé pour le *process mining*, serait par exemple un listing de patients (évènements) dont les interventions représenteraient les activités et leur enregistrements l'horodatage. La problématique pour appliquer le *process mining* aux données issues d'un match de football est de définir ce que seraient un évènement et une activité (l'horodatage étant naturellement l'enregistrement du temps d'une action observée pendant le match). Ce découpage est une question nécessitant une attention particulière : un découpage trop long engendre une perte de granularité, risquant ainsi d'occulter des nuances significatives dans le déroulement du jeu. D'autre part, un découpage excessivement court peut conduire à une fragmentation excessive des données, rendant difficile la capture des schémas de jeu plus larges et des dynamiques stratégiques. Ainsi, la détermination d'une unité de temps appropriée pour le découpage des événements demeure un enjeu essentiel dans l'analyse statistique du football. Plusieurs découpages ont été réalisés dans notre travail : (1) Découpage chronologique : on considère qu'un événement est délimité par un intervalle de temps, tel que, par exemple, toutes les 5 minutes. (2) Découpage post-activité : on considère qu'un événement est composé des X activités précédant un évènement spécifique, tel qu'un but. Si l'analyse se focalise sur un évènement spécifique, c'est plutôt le second découpage qui sera utilisé.

Dans les exemples qui sont traités par la suite, des représentations graphiques différentes sont proposées pour mettre en évidence et caractériser les types d'actions qui suivent ou précèdent un évènement spécifique. Ces représentations sont obtenues par des analyses de graphes ou des analyses exploratoires classiques sur les évènements et leurs durées respectives.

# 3 Comparaison de deux matchs pour l'équipe d'Espagne

## 3.1 Process map des actions précédant un but

Nous rapportons ici, par souci d'analyse rapide et d'interprétation pour un lecteur non familiarisé au football, un découpage post-activités composé des deux activités précédant un but. Pour le match d'ouverture 3-0, seulement deux buts ont été étudiés, l'autre but étant marqué par le Costa Rica contre son camp en faveur de l'Espagne. Pour le match terminal un seul but a été marqué en faveur de l'Espagne. Par conséquent, l'objectif de l'analyse ici présentée n'est pas d'inférer le déroulement d'un match futur mais plutôt de comparer deux matchs dont le niveau de l'équipe adversaire diffère fortement. Nous rappelons ici que la *process map* représente l'ensemble des séquences menant à un but sous forme d'un *Direct Follower Graph*. Nos résultats ont montré qu'il n'y avait pas de différence significative entre les séquences d'actions des deux matchs.

Lors du match d'ouverture, la complexité de la *process map* de gauche de la figure 1


Figure 1: *Process map* illustrant la séquence des deux dernières actions menant à un tir réussi pour l'Espagne lors de son premier match (à gauche) et de la finale (à droite).

indique plusieurs opportunités pour marquer un but. Deux sous-ensembles d'actions principaux conduisant à un but se dégagent. Pour la combinaison Dribble-Carry-Shot, il s'agit d'une opportunité de tir direct, comme cela a été observé dans d'autres matchs. La combinaison Ball Recovery-Shot, représente plusieurs tentatives précédant un tir réussi. Plus précisément, 16,67% des activités répertoriées impliquent des tentatives de reprise de possession du ballon par une joueuse adverse. Dans tous les cas où cette tentative a abouti, elle est suivie d'un contrôle du ballon au niveau du pied de la joueuse en mouvement, ce qui conduit systématiquement à un tir. En outre, un cas isolé montre qu'un tir sur trois est suivi d'une tentative de reprise de possession, ce qui indique que, lorsqu'une joueuse adverse intercepte le ballon après un tir dans la zone adverse, cela conduit toujours à un autre tir. En revanche, le *process map* de droite est linéaire en raison de l'unique but marqué lors de la finale. Le schéma de jeu gagnant, qui est très simple, indique qu'une passe bien reçue a été suivie par un tir réussi. Cette comparaison met en évidence :

- une diminution des occasions de but entre les deux matchs, en raison du niveau de l'équipe adverse.
- la variété des schémas d'actions dans le premier match qui reflète la diversité des situations de but potentielles, alors que les schémas d'actions (en incluant les buts manqués) sont quasi identiques pour le dernier match. Cela peut s'expliquer par une efficacité du jeu dans un contexte de pression intense et de compétition décisive.

Ces résultats, évidents du point de vue du football, illustrent l'intérêt d'une démarche de *process mining* qui permet d'étudier des séquences de jeu menant à des évènements spécifiques au cours d'un match.

## 3.2 Actions menant le ballon hors du terrain

Nos analyses sur les sorties de terrain (Fig. 2) ont montré qu'une plus grande variété de types d'actions conduit le ballon hors des limites du terrain pendant la finale. Les passes représentent 17% des actions précédant les sorties de balle dans le match terminal contre 11% dans le match d'ouverture. Par ailleurs, la stratégie de l'équipe diffère entre l'Angleterre qui incite l'Espagne à commettre davantage de fautes, devenant la deuxième

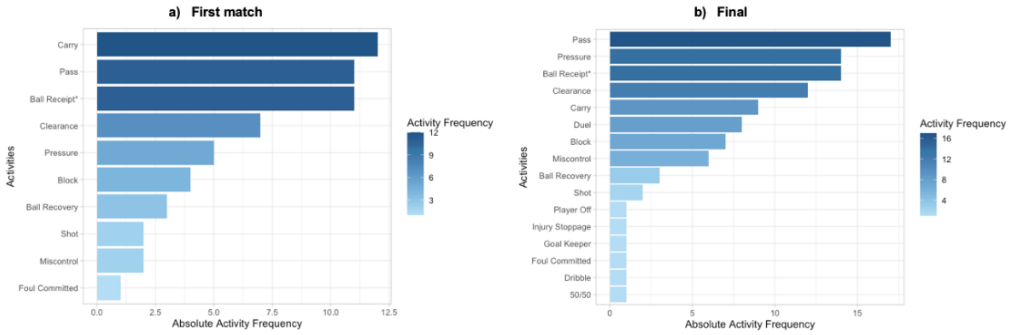



Figure 2: Bar Chart illustrant la distribution de fréquence des actions de l'Espagne menant à une sortie de terrain lors du premier match de l'Espagne (à gauche) et de la finale (à droite).

cause des sorties de ballons, soit 14% des actions précédentes, contre seulement 5% lors du match contre le Costa Rica. Cette évolution suggère une intensification de la défense anglaise, mettant en évidence une pression accrue sur l'équipe espagnole et un ajustement stratégique de sa part pour contrer cette pression accrue.

## 3.3 Actions suivant la remise en jeu par la gardienne de but

L'analyse comparative des deux graphiques de la figure 3 se concentre sur les deux actions qui suivent un coup de pied par la gardienne. On s'intéresse donc ici aux types d'évènements ainsi qu'à leur durée. Cette analyse révèle des différences significatives entre le premier match et la finale.

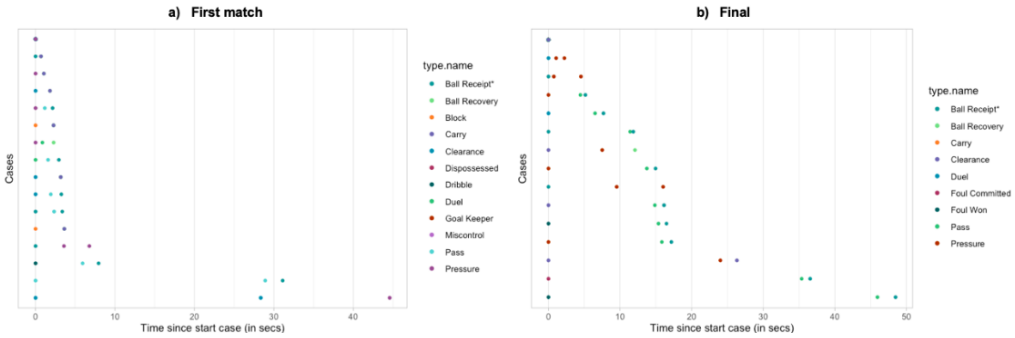



Figure 3: Graphique illustrant les quatre types d'actions consécutives à un coup de pied de la gardienne espagnole lors du premier match (à gauche) et de la finale (à droite). Chaque case est un évènement spécifique.

Contrairement à d'autres études d'évènements, les durées d'évènements sont similaires pour les deux matches (globalement inférieurs à 15 sec). Ici 75% des évènements des actions se produisent dans les 5 secondes qui suivent le coup de pied de la gardienne, avec une extension jusqu'à 40 secondes pour les cas les plus longs. Les outliers du graphique associé au premier match (4 points les plus à droite sur le graphique de gauche) correspondent aux premières actions se produisant près de 30 secondes après le début de la remise en jeu. A l'inverse, dans le match terminal (à droite) environ un tiers des évènements voient leur première action se produire dans les quinze secondes, tandis que les deux tiers restants ont leur première action entre 35 et 45 secondes après le coup de pied. Cette analyse met en évidence des dynamiques temporelles distinctes entre les deux matches, suggérant des variations dans les schémas de jeu et l'évolution des situations après les remises en jeu de la gardienne, notamment pour la finale suggérant un changement de stratégie à des moments clés spécifiques du match.

# 4 Conclusion

En conclusion, l'analyse des performances de l'Espagne lors de la Coupe du monde de football féminin à travers l'exploration des process, notamment entre les matchs initial et final, permet de contribuer à comprendre le parcours de l'équipe. L'examen des résultats des tirs met en évidence les variations dans les stratégies offensives, en soulignant la réaction face à la défense ainsi que les actions de la gardienne de but. La finale est marquée par une défense agressive de l'Angleterre. En réaction, l'équipe d'Espagne a mis en place des ajustements tactiques et des schémas de jeu potentiellement plus prévisibles en finale par rapport au jeu variable contre le Costa Rica. Cette dynamique contrastée du premier match contre le Costa Rica, caractérisée par des tentatives de tir différées, met en évidence la capacité d'adaptation de l'Espagne. La diminution des occasions de but en finale se présente sous la forme d'un *processus simplifié*. Le changement des temps d'évènements, en particulier après les coups de pied de but, suggère des ajustements stratégiques ou des moments clés dans le match final. Malgré les variations dans le temps de possession, les joueuses espagnoles adaptent très rapidement leurs positions. Cependant, l'interprétation des données reste complexe, soulignant la nécessité de prendre en compte de multiples facteurs contribuant à la dynamique du match. L'analyse des passes espagnoles tout au long de la compétition indique une amélioration générale de la précision, avec une diminution significative des passes hors limites. En conclusion, cet article fournit des informations sur les performances de l'Espagne à travers une analyse statistique, illustrant sa capacité d'adaptation par des ajustements tactiques. Notre travail montre également que l'approche par l'exploration des processus permet d'analyser des séries d'actions sous forme de séquence, et se révèle utile pour comprendre les complexités inhérentes liées au football. Comme travaux futurs nous envisageons de détecter des similarités entre différents couples équipe/match à l'aide d'une classification de *process*.

# References


[1] van der Aalst, W.: Data Science in Action, pp. 3–23. Springer Berlin Heidelberg, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4˙1https://doi.org/\detokenize{10.1007/978-3-662-49851-4_1}

[2] Janssenswillen, G., Depaire, B., Swennen, M., Jans, M.J., Vanhoof, K.: bupaR: Enabling reproducible business process analysis. Knowledge-Based Systems **163**, 1857 (2019). https://doi.org/10.1016/j.knosys.2018.10.018https://doi.org/\detokenize{10.1016/j.knosys.2018.10.018}

[3] Peter O'Donoghue, Katerina Papadimitriou, V.G., Haralambis, K.: Statistical methods in performance analysis: an example from international soccer. International Journal of Performance Analysis in Sport **12**(1), 144–155 (2012). https://doi.org/10.1080/24748668.2012.11868590https://doi.org/\detokenize{10.1080/24748668.2012.11868590}

[4] Rico-González, M., Pino-Ortega, J., Praça, G.M., Clemente, F.M.: Practical applications for designing soccer' training tasks from multivariate data analysis: A systematic review emphasizing tactical training. Perceptual and Motor Skills **129**(3), 892–931 (2022). https://doi.org/10.1177/00315125211073404https://doi.org/\detokenize{10.1177/00315125211073404}, pMID: 35084256

# Apprentissage automatique pour l'identification des caractéristiques de jeu d'une équipe victorieuse au Rugby à XV

Arnaud Odet [1,2], Thomas Béchard [2], Pierre Moretto [2], Sébastien Déjean [1] & Cristian Pasquaretta [2]

[1] *Institut de Mathématiques de Toulouse, UMR 5219 Université de Toulouse et CNRS {arnaud.odet, sebastien.dejean}@math.univ-toulouse.fr*

[2] *Centre de Recherches sur la Cognition Animale (CRCA), Centre de Biologie Integrative (CBI), CNRS, Université Paul Sabatier, Toulouse III, France {thomas.bechard, pierre.moretto ,cristian.pasquaretta}@univ-tlse3.fr*

**Résumé.** La prédiction de résultats de rencontres sportives est un sujet qui connaît un intérêt croissant depuis quelques années, bien aidé par la démocratisation de techniques d'apprentissage automatique. Néanmoins, l'utilisation des modèles prédictifs à des fins d'amélioration de la performance tactique des équipes dans les sports collectifs reste, à notre connaissance, limitée. C'est avec l'objectif de contribuer à ce sujet, et à travers un cas d'étude sur le Rugby à XV, que nous proposons une méthodologie combinant apprentissage automatique et techniques d'explicabilité des algorithmes. Le présent travail se déroule en deux temps : tout d'abord, nous avons établi un modèle de prédiction de résultat sur la base d'indicateurs de performances observés au cours des matchs, puis nous avons appliqué aux prédictions de ce modèle une analyse basée sur les *SHAP values*. Les résultats permettent en se plaçant (i) dans une perspective locale de proposer aux staffs techniques des analyses diagnostiques au sujet des matchs passés et (ii) dans une perspective globale de définir les indicateurs de performances déterminant l'issue d'un match. Nos résultats soulignent l'importance des jeux au pied, des plaquages, et des franchissements.

**Mots-clés.** Analyse diagnostique, *SHAP (SHapley Additive exPlanations) values*

**Abstract.** Predicting the outcomes of sports matches has garnered increasing interest in recent years, fueled by the development and popularization of machine learning techniques. However, the utilization of newly developed predictive models for the purpose of enhancing tactical team performance in team sports remains, to our knowledge, limited. With the aim of contributing to this field, and through a case study on Rugby Union, we suggest a methodology that combines machine learning and algorithm explainability techniques. This study unfolds in two phases : first, we sought to identify the type of modeling that best suits our data, establishing a prediction model based on performance indicators observed during games. Subsequently, we applied an analysis based on SHAP values to the predictions of this model. Our findings serve two primary purposes : (i) from a local perspective, providing technical staff with diagnostic analyses regarding past matches, and (ii) from a global standpoint, identifying which performance indicators mostly determine the match outcomes. Our results emphasize the importance of kicking games, tackling, and breakthroughs.

**Keywords.** Diagnosis analysis, SHAP (SHapley Additive exPlanations) values

# 1 Introduction

L'analyse de données appliquée à la prédiction des résultats des matchs dans le domaine du sport collectif connaît un engouement croissant ces dernières années en Europe.

Bien que son efficacité soit aujourd'hui largement reconnue (revue par Horvat et Job, 2020), la littérature scientifique demeure divisée quant au choix des modèles prédictifs les plus adaptés (Bunker et Susnjak, 2022). Les méthodes utilisées peuvent être des modèles de type *black-box* (réseaux de neurones, support vector machine, méthodes d'ensembles) ou *white-box* (méthodes linéaires). Cependant, elles sont rarement comparées et motivées.

Les prédictions sont en majorité abordées comme un problème de classification binaire (gagner ou non un match). Les prédicteurs utilisés pour résoudre ce problème de classification peuvent être de deux types : tirés des résultats de matchs passés (par exemple nombre de points marqués, classement avant la rencontre, etc. voir Lampis et al., 2023) ou des indicateurs de performances au sein des matchs (par exemple nombre de passes, nombre de tirs, etc. voir Miljkovic et al., 2010). La différence fondamentale entre ces deux types de prédicteurs est que les résultats des matches précédents n'apportent aucune information sur la stratégie d'une équipe contrairement aux indicateurs de performance qui peuvent quantifier la pertinence des choix tactiques et techniques.

A notre connaissance, malgré les avancées significatives dans la prédiction de résultats de matchs sportifs, l'utilisation de telles analyses pour guider les décisions des staffs techniques reste encore marginale (e.g. Shin et Gasparyan, 2014).

Dans cette optique, notre objectif est double. Nous cherchons d'abord à comparer différents types de modélisation pour déterminer la plus pertinente à prédire l'issue d'un match. A des fins d'exploitation tactique, nous proposons ensuite d'adopter une approche *model agnostic* qui permet l'intégration des résultats obtenus dans les analyses diagnostiques. Cette méthodologie est ici appliquée au rugby à XV.

# 2 Les données

## 2.1 Description des données

Les données ont été obtenues à partir du site aiasports.fr[1] pour les saisons 2021/2022 et 2022/2023 des championnats suivants : Premiership, Top14, ProD2, Super Rugby et United Championship, et contient des matchs de saison régulière. Les données ont été extraites de fichiers d'analyse de matchs, puis agrégées pour former un jeu de données contenant 1567 matchs.

Le jeu de données comprend de nombreuses variables, disponibles pour l'équipe locale (*home*) et extérieure (*away*). Ces variables peuvent être un nombre (e.g. nombre de rucks réussis) ou un pourcentage (e.g. pourcentage de rucks réussis). Certaines de ces variables étant

[1] https ://www.aiasports.fr/

des combinaisons linéaires d'autres variables (par exemple, les totaux sont la somme des sous totaux), ou fortement colinéaires à d'autres variables (les pourcentage se calculent à partir de la variable d'intérêt et des autres sous-totaux), elles seront exclues de l'analyse. De plus, les caractéristiques liées au score (nombre d'essais, pénalités, etc) ne sont pas considérées dans l'analyse dans la mesure ou une combinaison linéaire de ces indicateurs suffit à déterminer le résultat du match, et d'autres variables ont été ignorées, car incomplètes[2].

Nous avons donc retenu un jeu de données de 40 variables. Ce sont les possessions (8 variables : nombre, durée, position dans l'air de jeu, occupation), les touches (3 variables : nombre, indicateur de qualité), les mêlées (3 variables : indicateurs de qualité), les jeux au pied (6 variables : taux de réussite, dégagement, terrain, pression, hors terrain, contest), les rucks (7 variables : durée, position dans l'aire de jeux, réussite), les passes (2 variables : simples, après contact), les franchissements (5 variables : position dans l'air de jeu, nombre de défenseurs battus), les plaquages (4 variables : indicateurs de qualité) et les cartons (2 variables : jaune et rouge).

La position sur le terrain (d'une longueur d'environ 100m) se décompose en 4 zones : lorsque l'équipe est dans ses 22 mètres, entre ses 22 mètres et la ligne médiane, entre la ligne médiane et les 22 mètres adverses, dans les 22 mètres adverses.

Les touches, les mêlées et les plaquages sont aussi repartis en 3 groupes en indiquant leur qualité : positifs, neutres, et négatifs. Ainsi, une touche positive est une touche non contestée lors de laquelle l'alignement adverse ne saute pas ou pas en miroir pas clair de l'alignement de l'équipe effectuant la remise en jeu. Une touche neutre est contestée mais l'équipe effectuant la remise en jeu conserve le contrôle du ballon, et une touche négative est une touche perdue. Pour les mêlées, la distinction est similaire : une mêlée positive est une mêlée remportée par l'équipe introduisant la balle, sans que la sortie de balle soit directement contestée. Une mêlée neutre est une mêlée dont la sortie de balle est directement contestée ou injouable (sanctionnée comme telle par l'arbitre, et re-jouée avec la même équipe introduisant la balle). Enfin, une mêlée négative est une mêlée où l'équipe introduisant la balle la perd. Pour les plaquages, un plaquage positif est un plaquage lors duquel la défense gagne la ligne d'avantage obligeant l'attaque à reculer. Un plaquage neutre est un plaquage lors duquel la ligne d'avantage est conservée et un plaquage négatif est un plaquage lors duquel la ligne d'avantage est gagnée par l'attaque, qui avance. Considérant que les plaquages positifs, neutres et négatifs sont réussis, nous disposons également du sous total de plaquages réussis. La qualification de ces données est effectuée par le site aiasports.fr et n'a demandé aucune intervention de notre part.

Les durées des possessions s'expriment en secondes et sont classées par catégories : de 0 à 30 secondes, de 30 à 60 secondes, plus de 60 secondes. Les durées des rucks sont également exprimées en secondes et classées par catégories : entre 0 et 3 secondes, entre 3 et 6 secondes, plus de 6 secondes.

[2]Présentes uniquement sur les matchs les plus récents.

## 2.2 Création de variables

Dans le but de tenir compte de la différence de niveau intrinsèque entre deux équipes, nous avons créé une variable "points ELO" (Elo, 1978), avec un facteur $K$ fixé à 20, en appliquant les valeurs utilisées par la fédération internationale d'échecs (FIDE)[3]. Nous avons considéré que la situation de ligues ouvertes où les équipes sont regroupées par niveau ne correspondait pas à une situation de nouvelle arrivée dans le système de classement (cas où $K = 40$ pour la FIDE) et que le faible nombre d'équipes professionnelles (comparé au nombre de joueurs d'échecs présents dans le classement FIDE) justifiait de ne pas établir un seuil au-delà duquel l'échange de points est plus faible (cas où $K = 10$ pour la FIDE). Selon Lampis et al. (2023), la détermination de cette variable $K$ peut faire l'objet d'une recherche approfondie.

Les points ELO sont initialisés au début de chaque saison afin (i) de tenir compte de la variabilité inter-saisons des effectifs, et (ii) d'intégrer les montées vs descentes entre les différentes divisions.

Considérant que les premiers matchs d'une saison sont d'une part des matchs où les points ELO ne reflètent pas encore le niveau des équipes, et d'autre part servent à de nombreuses mises au point et choix technico-tactiques, nous avons évincé les données des 4 premières rencontres.

Par ailleurs, nous avons classé les matchs suivant qu'ils étaient remportés par l'équipe locale (1) ou pas (0). Une limite de ce choix est la considération de 45 matchs nuls (2.9%), qui sont donc classés dans la même catégorie qu'une défaite de l'équipe locale. Cette approche est justifiable par l'existence du *home advantage* dans le sport (Courneya et Carron, 1992).

# 3 Méthode

## 3.1 Description des modèles choisis

Nous avons utilisé différents types de modèles parmi les plus employés (y compris dans les travaux liés au sport, voir Bunker et Susnjak, 2022) : des modèles linéaires (régression logistique, analyse discriminante par moindres carrés partiels ou PLS-DA), des méthodes d'ensembles (Random Forest, Adaptative Boost classifier, XGBoost Classifier), de Support Vector Machine (avec des kernels linéaires et RBF), un *k-Nearest-Neighbors*, et deux modèles de type *Artificial Neural Network*, qui diffèrent par leurs architectures. Nous avons également utilisé deux modèles de bases afin d'évaluer le gain de capacité de prédiction apporté par notre analyse : le modèle *Baseline Home*, qui prédit systématiquement une victoire de l'équipe locale (67% des matchs de notre jeu de données), et le modèle *Baseline ELO*, qui consiste en une régression logistique avec pour seul prédicteur la différence de points ELO entre les deux équipes avant le début de la rencontre.

[3]https ://handbook.fide.com/chapter/B022024

## 3.2 Métriques de comparaison

Nous avons choisi d'utiliser le modèle le plus performant sur la base des *accuracy* respectives. Nous avons également affiché le score F1 afin de rendre compte de la sensibilité du modèle aux faux positifs et faux négatifs.

L' *accuracy* est une métrique communément utilisée pour définir la qualité d'un modèle de classification. Elle se définit comme étant la fraction des individus correctement classés. Le score $F1$ est la moyenne harmonique de la *precision* et du *recall*. Ces métriques se calculent comme suit :

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad F1 = \frac{2TP}{2TP + FP + FN}$$

où $TP$ et $FP$ sont respectivement le nombre de vrais (faux) positifs (*True Positive / False Positive*) et $TN$ et $FN$ sont respectivement le nombre de vrais (faux) négatifs (*True Negative / False Negative*)

## 3.3 Choix des données d'entraînements et de prédiction

Nous avons gardé les 4 dernières journées de chaque saison pour l'évaluation. Le choix des 4 dernières journées permet de garder environ 20% des données dans le *test set*. Ainsi, comme énoncé en 2.2, les 4 premières journées de chaque championnat ne sont pas utilisés (280 matchs), le modèle est entraîné sur les matchs des journées suivantes (1014 matchs), et évalué sur les 4 dernières journées (273 matchs[4]).

# 4 Résultats

## 4.1 Comparaison des modèles

Les résultats de la comparaison des modèles sont présentés dans la Table 1. On constate que le modèle Support Vector Classifier (Kernel RBF) obtient les meilleurs scores, tant en termes d'*accuracy* que de *F1-score*.

Il convient de noter que si les modèles linéaires (régression logistique et PLS-DA) performent moins que le modèle Support Vector Classifier (Kernel RBF), ils présentent l'avantage d'être plus facilement explicables.

Pour nos modèles de références, et contrairement aux attentes, le modèle naïf prédisant systématiquement une victoire de l'équipe locale obtient la même *accuracy* que le modèle de référence utilisant la différence de score ELO précédent la rencontre. Il présente également un score F1 supérieur.

[4]La différence du nombre de matchs entre les 4 premières et les 4 dernières journées s'explique par l'abandon des équipes Wasps et Worcester Warriors en cours de saison en Premiership 2022/23.

Table 1 : Tableau récapitulatif des Accuracy et des F1-score des différents modèles

| Model | Accuracy | F1-score |
|---|---|---|
| Support Vector Classifier - RBF kernel | 0.8168 | 0.8744 |
| Support Vector Classifier - Linear kernel | 0.7912 | 0.8496 |
| Partial Least Squares - Discriminant Analysis | 0.7875 | 0.8490 |
| Logistic Regression | 0.7839 | 0.8468 |
| ANN - 1 hidden layer | 0.7839 | 0.8468 |
| ANN - 3 hidden layers | 0.7839 | 0.8468 |
| Ada Boost | 0.7766 | 0.8373 |
| Random Forest | 0.7253 | 0.8219 |
| XGB Classifier | 0.7253 | 0.8072 |
| k - Nearest Neighbors Classifier | 0.6996 | 0.7960 |
| Baseline Home | 0.6923 | 0.8182 |
| Baseline ELO | 0.6923 | 0.8000 |

## 4.2 Analyse diagnostique

Afin d'expliquer un match donné, et fournir à l'encadrement technique d'une équipe les raisons susceptibles d'avoir conduit au résultat final, nous proposons d'analyser les prédictions du modèle avec la méthodologie développée par Lundberg et Lee (2017), introduisant les *SHAP (SHapley Additive exPlanations) values*. Cette méthodologie permet de mesurer l'impact des différentes variables sur la prédiction. Elle se base sur les valeurs de Shapley (Shapley, 1953), définies comme suit :

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

où $\phi_i$ est la valeur de Shapley associé au joueur $i$, $S$ est une sous-coalition de $N$ ne contenant pas $i$ et $v(S)$ est la valeur de la coalition.

En théorie des jeux, une coalition $S$ est un groupe de joueurs obtenant dans un jeu collaboratif la valeur $v(S)$ et une sous-coalition de $S$ est un sous ensemble de joueurs $j \in S$. La coalition $N$ désigne l'ensemble des joueurs. L'objectif des valeurs de Shapley est d'attribuer à chaque participant d'une coalition une fraction de la valeur dégagée par la coalition proportionnelle à l'apport de ce joueur à la coalition. En apprentissage automatique, chaque variable (prédicteur) est considérée comme un joueur, la coalition analysée est l'ensemble des variables utilisées par un modèle, et enfin la valeur dégagée par la coalition est la valeur prédite par le modèle. Ainsi, l'objectif est de déterminer quelles sont les variables qui contribuent le plus à la prédiction observée.

Appliquée à l'échelle locale, cette méthode permet de montrer quelles sont les variables qui ont conduit à l'élaboration de la prédiction (cf Figure 1, résultat de cette analyse pour un match de Top 14 de la saison 2022/23). Dans l'exemple présenté dans la Figure 1, le modèle prédit une victoire de l'équipe locale (avec une probabilité de 91.2%) due principa-

lement aux variables suivantes : le nombre de franchissements dans les 22 mètres adverses de l'équipe locale a contribué positivement à la prédiction d'une victoire de l'équipe locale à hauteur de +5%, le nombre de jeux au pied de dégagement de l'équipe visiteuse a contribué négativement à cette même prédiction à hauteur de -4%. Le nombre de touches perdues par l'équipe visiteuse, de jeux au pied de terrain de l'équipe locale et son nombre de possessions dans les 22m adverses ont chacune contribué positivement à hauteur de +4%. Sachant que le match en question a été gagné par l'équipe locale, le staff technique de l'équipe visiteuse pourrait par exemple décider d'une utilisation différente des jeux au pied de terrain puisque cette variable a contribué à hauteur de +3% à la probabilité de victoire de l'équipe locale.

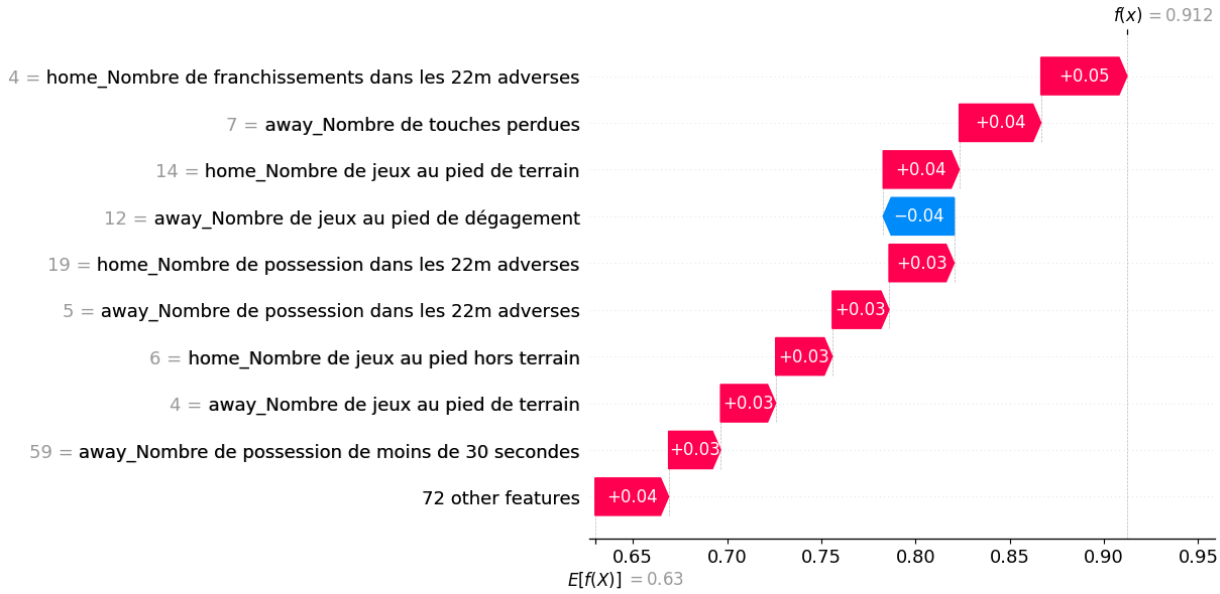



Figure 1 : Illustration de l'explication d'une prédiction locale pour un match de Top 14 lors de la saison 2022/23. L'axe des abscisses représente la probabilité que l'équipe locale remporte le match selon le modèle, et les valeurs présentées sur l'axe des ordonnées sont les valeurs des variables correspondantes. Une variable dont la valeur sur le match en question pousse le modèle à prédire un 1 (victoire de l'équipe domicile) est indiquée en rouge, et une variable dont la valeur pousse le modèle à prédire un 0 est indiquée en bleu. Les valeurs indiquées sur les flèches sont l'impact estimé de la valeur du prédicteur en pourcentage sur la probabilité de victoire de l'équipe locale.

Appliquée à l'échelle globale, elle permet de mettre en lumière les variables qui influent le plus sur l'issue d'un match de rugby sur notre jeu de test, comme l'illustre la Figure 2. À titre d'exemple, le nombre de passes simples de l'équipe locale semble être un facteur qui est négativement corrélé avec la probabilité de victoire de cette même équipe. En effet, les points où cette variable prend une valeur élevée se situent à gauche de l'axe central, indiquant que cette variable pousse la prédiction vers le 0. Les jeux au pied de terrain semblent à l'inverse être positivement corrélés avec les chances de victoires pour l'équipe qui en fait un nombre élevé. Cette observation est illustrée par les points rouges à gauche (respectivement droite) de l'axe central pour l'équipe visiteuse (respectivement locale). La variable ayant une plus faible concentration de points proches de l'axe central est donc la variable qui a le plus d'impact

sur les prédictions du modèle. Ici, la différence de points ELO précédant la rencontre semble être la variable impactant le plus le résultat de la rencontre.

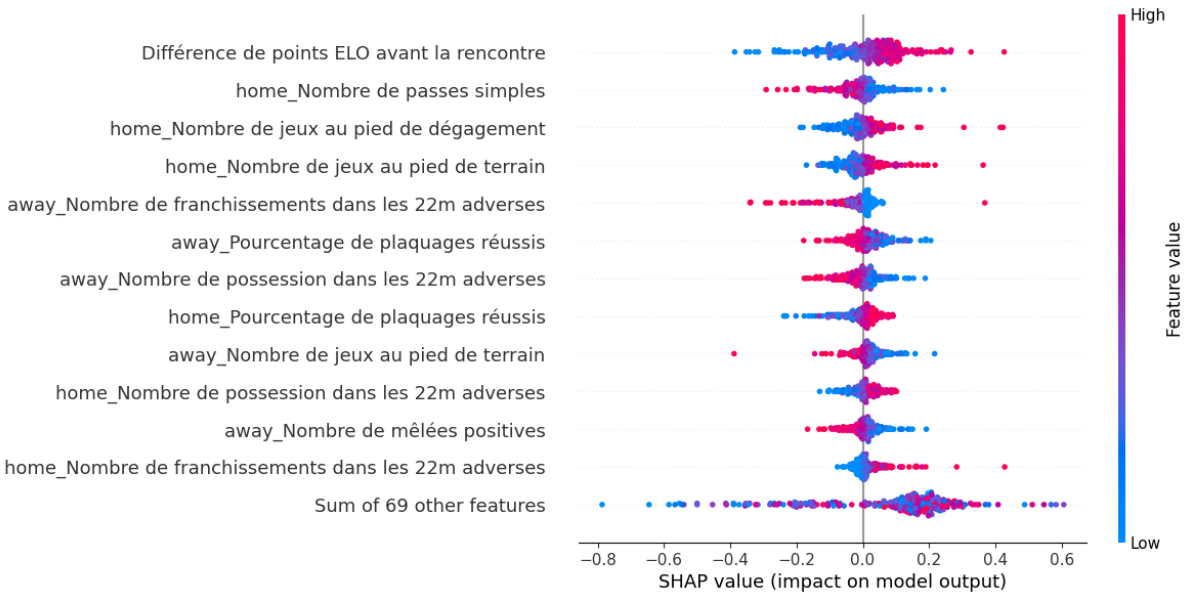



Figure 2 : Illustration de l'explication à l'échelle globale du modèle Support Vector Classifier (Kernel RBF). La coloration des points indique s'ils correspondent à une valeur élevée (rouge) ou faible (bleus) pour la variable en question, et la position sur l'axe horizontal représente l'impact associé aux points correspondants. Vers la gauche le facteur prédit un 0 et vers la droite un 1. Ainsi, plus une variable compte de points loin de l'axe central, plus son pouvoir d'influence sur le dénouement d'une rencontre est grand.

En conséquence, une hiérarchisation de la magnitude de l'impact des variables peut être proposée en fonction de l'écart type de leurs *SHAP values* (comme présenté sur la Figure 3). La différence de points ELO n'apparaît pas sur la Figure puisque c'est une variable commune aux deux équipes, mais elle reste la variable dont l'écart type des *SHAP values* est le plus important (0.11). Il apparaît alors que les variables les plus déterminantes sur notre jeu de données sont les franchissements dans les 22m adverses, le pourcentage de plaquages réussis, et le nombre de jeux au pied de terrain.

# 5 Discussion et conclusion

Dans l'étude actuelle, nous avons proposé une méthode permettant d'exploiter des indicateurs de performance pour fournir aux staffs techniques (i) une analyse globale des indicateurs de performance les plus pertinents au rugby et (ii) à niveau local un diagnostic des matchs écoulés.

Sur nos données, le Support Vector Classifier (Kernel RBF) affiche une *accuracy* de 81.68% sur le jeu de données de test, et révèle des bons prédicteurs de performance au rugby. En

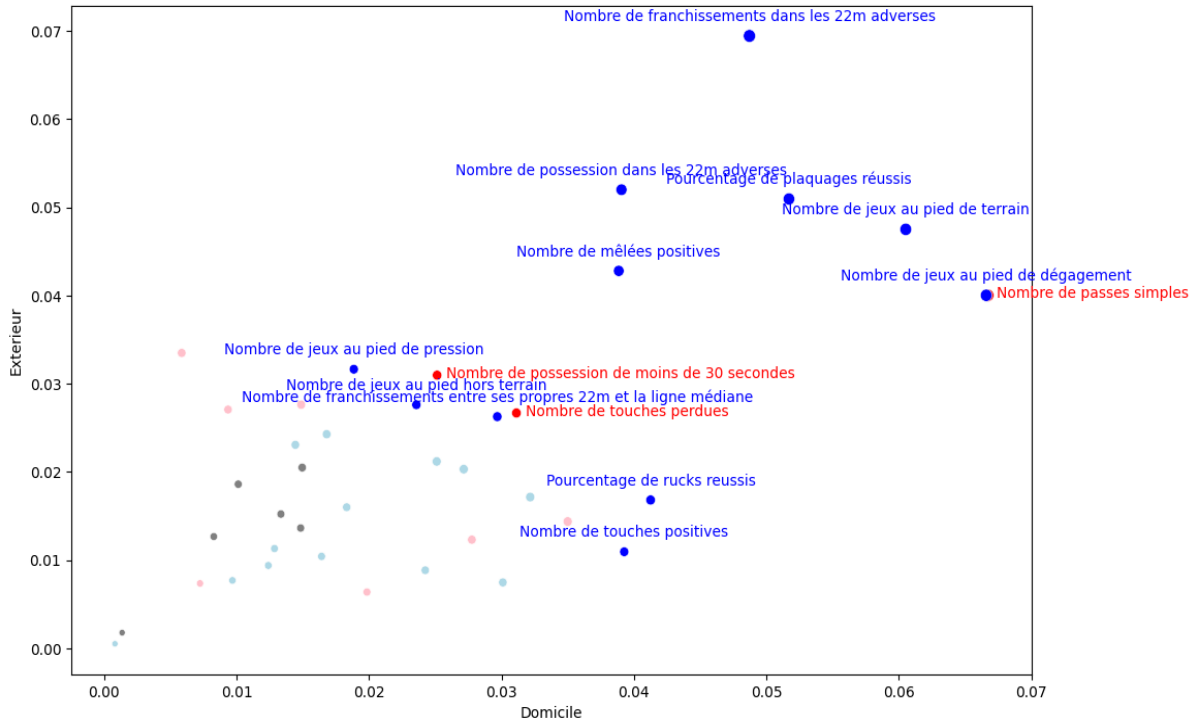



Figure 3 : Ecarts types des *SHAP values* des variables selon le modèle Support Vector Classifier (Kernel RBF). L'axe horizontal et l'axe vertical représentent respectivement l'importance d'une variable pour l'équipe locale et visiteuse. La coloration des points indique une corrélation positive (bleue) ou négative (rouge) avec les *SHAP values*, et, par conséquent, que les équipes ont intérêt de maximiser (ou minimiser) pour augmenter leurs chances de victoires. Les points gris représentent des variables dont l'impact n'est pas clairement identifié. Afin de préserver la lisibilité de la figure, seules les variables les plus significatives ($\frac{\sigma_{home}+\sigma_{away}}{2} > 0.25$ où $\sigma$ représente l'écart type), sont nommées, les autres apparaissent dans une couleur plus claire.

particulier la probabilité de gagner un match est positivement influencée par (i) nombre de franchissements dans les 22m adverses, (ii) un plus haut pourcentage de plaquages réussis, et (iii) un nombre supérieur de jeux au pied de terrain. Il est intéressant de noter que le nombre de passes simples apparaît comme négativement corrélé avec la probabilité de victoire au rugby à XV.

Bien que d'importance globale, ces prédicteurs peuvent influencer différemment le résultat lorsqu'ils sont observés au niveau d'un seul match (e.g. Figure 1). Notre méthode permet une vision multi-échelle et peut apporter un nouvel éclairage sur la stratégie à adopter dans la préparation des matchs.

Si notre travail n'a aucunement la prétention de se substituer à l'expérience et l'expertise d'un staff technique, nous pensons qu'il peut apporter un complément d'information, et des éléments d'aide à la décision, à travers l'analyse des points faibles et points forts de leur propre équipe ou des adversaires. Par ailleurs, l'utilisation de techniques d'apprentissage automatique peut permettre d'analyser un nombre de matchs qui serait impossible à traiter

pour un staff technique.

Notre méthode peut encore être améliorée car elle est basée sur un volume relativement restreint de données, ce qui peut avoir un impact sur la robustesse de certaines conclusions. La quantité de variables à disposition et les choix faits pour éviter de sélectionner des variables colinéaires ont également un impact sur la performance des différents modèles. Ensuite, une difficulté réside dans le fait que simuler l'absence d'une variable n'est pas trivial, or l'utilisation des *SHAP values* repose sur de telles simulations. En effet, elle suppose d'attribuer une valeur issue d'un *background* à la variable dont l'absence est simulée, et le choix de ce dernier influence les résultats (voir Albini et al., 2022). Par ailleurs, pour compléter le présent travail, il conviendrait de tenir compte d'une certaine dépendance entre les variables et de modéliser le report modal : pour reprendre l'exemple développé en Figure 1, si l'équipe visiteuse décidait d'utiliser davantage les jeux au pied de terrain, l'équipe aurait nécessairement moins de passes et/ou de rucks et/ou de franchissements.

Notons enfin que la méthode présentée, et illustrée ici à travers un cas d'étude sur le rugby à XV, est transposable dans d'autres sports.

## Remerciements

Nous remercions la société AIA SPORTS qui nous a permis d'utiliser les données mentionnées.

## Références

Albini, E., Long, J., Dervovic, D., & Magazzeni, D. (2022). Counterfactual Shapley Additive Explanations. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Bunker, R., & Susnjak, T. (2022). The Application of Machine Learning Techniques for Predicting Match Results in Team Sport : A Review. *Journal of Artificial Intelligence Research*, *73*.

Courneya, K., & Carron, A. (1992). The Home Advantage In Sport Competitions : A Literature Review. *Journal of Sport & Exercise Psychology*, *14*.

Elo, A. (1978). The rating of chessplayers, past and present.

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction : A review. *WIREs Data Mining and Knowledge Discovery*, *10*(5).

Lampis, T., Ioannis, N., Vasilios, V., & Stavrianna, D. (2023). Predictions of european basketball match results with machine learning algorithms. *Journal of Sports Analytics*, (Preprint).

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Miljkovic, D., Gajic, L., Kovacevic, A., & Konjovic, Z. (2010). The use of data mining for basketball matches outcomes prediction. *IEEE 8th International Symposium on Intelligent Systems and Informatics*.

Shapley, L. (1953). A value for n-person games.

Shin, J., & Gasparyan, R. (2014). A novel way to Soccer Match Prediction.

# Understanding the Dynamics of Women's Football through Clustering

Yvenn Amara-Ouali [1,2]

[1] *University Paris-Saclay, LMO, Orsay, FRANCE,*
*yvenn.amara-ouali@universite-paris-saclay.fr*
[2] *EDF R&D, OSIRIS, Palaiseau, FRANCE*

**Résumé.** La montée en puissance du football féminin ces dernières années nécessite une compréhension plus approfondie de sa dynamique et des caractéristiques des joueuses. En utilisant des techniques statistiques avancées, cet article vise à explorer les subtilités du football féminin grâce à analyse par *clusters*. En analysant les données de performance des joueuses provenant de diverses compétitions, nous cherchons à identifier des types de joueuses distincts en fonction de leurs postes et de leurs styles de jeu. Nous utilisons une combinaison de modèles statistiques, d'algorithmes d'apprentissage automatique et de techniques d'apprentissage profond pour regrouper efficacement les joueuses. Nos résultats préliminaires montrent le potentiel de ces méthodes pour caractériser la diversité et la complexité du football féminin. À travers cette recherche, nous visons à fournir des informations pertinentes pour les entraîneurs et analystes afin d'améliorer le développement des joueuses, la tactique et le niveau global du sport.

**Mots-clés.** Football féminin, *clustering*, caractérisation statistique, apprentissage automatique, apprentissage profond.

**Abstract.** The rising prominence of women's football in recent years necessitates a deeper understanding of its dynamics and player characteristics. Leveraging advanced statistical techniques, this paper aims to explore the nuances of women's football through clustering analysis. By analyzing player performance data from various competitions, we seek to uncover distinct player types based on their positions and playing styles. We employ a combination of statistical models, machine learning algorithms, and deep learning techniques to cluster players effectively. Our preliminary results demonstrate the potential of these methods in characterizing the diversity and complexity of women's football. Through this research, we aim to provide valuable insights for coaches, analysts, and stakeholders to enhance player development, team strategies, and overall game performance.

**Keywords.** Women's football, clustering analysis, player characterization, statistical models, machine learning, deep learning.

## 1 Women's football on the rise

Women are certainly not new to the game of football, with one of the oldest games recorded taking place in 1881 in Scotland. However, in recent years the number of competitions,

professional female players and spectators has significantly increased. In the 2023 FIFA Women's World Cup, ticket sales targets were smashed as 1,978,274 fans watched the matches played in the tournament across ten stadiums [1]. As the number of competitions and players increase, the emergence of advanced statistics and data can help to better characterise and capture the subtleties of the game. This will be particularly relevant in the run up to and during the upcoming Olympics in Paris 2024.

In 2012, Lydia Nsekera, the President of the Burundi FA, became the first co-opted female member of the FIFA Executive Committee [2]. This comes in contrast with the discussion of the importance of women's football in the wider society by [3], "Through their participation they are destabilising notions about gender roles, as well as notions about the physical and natural difference between men and women, masculinity and femininity." As professional women's football grows in popularity, it is vital that data and statistics are used to improve the sport and to measure and highlight ways to increase and ensure equality between the men's and women's game.

## 2 Related Work

The field of Women's football research has been evolving significantly over the past two decades [2]. Multiple studies have been led to characterise female football players [4]. For instance, the study presented in [5] examined morphological differences, body composition, and lower extremity explosive strength among 20 female football players in the Serbian Super League, categorizing them by playing positions. Significant disparities in explosive power were noted between midfielders and attackers, but no notable variations in body composition emerged. This might indicate that physical abilities and advanced in-game statistics might be an alternative way to characterise women's football players.

The work we propose builds upon the analysis proposed at the 2023 StatsBomb conference with a work on clustering women's football players [6]. The authors used a PCA over a wide range of event features to then perform clustering using a Bayesian Gaussian Mixture Model (BGMM) including players with more than 1000 minutes game time. However, we've found that the clustering performed may have some limitations as it gathers all players regardless of their position throughout the game (see Figure 1). Our goal with this proposed paper is to correct this inherent bias by performing clustering of players for each player position on the pitch to actually derive features of players that are specific in a certain position played.

## 3 Methods

In this section we briefly introduce the methods that will be used to perform clustering using statistical models (e.g., Gaussian Mixtures), machine learning (e.g., k-means and DBSCAN) and deep learning (e.g, Autoencoders).

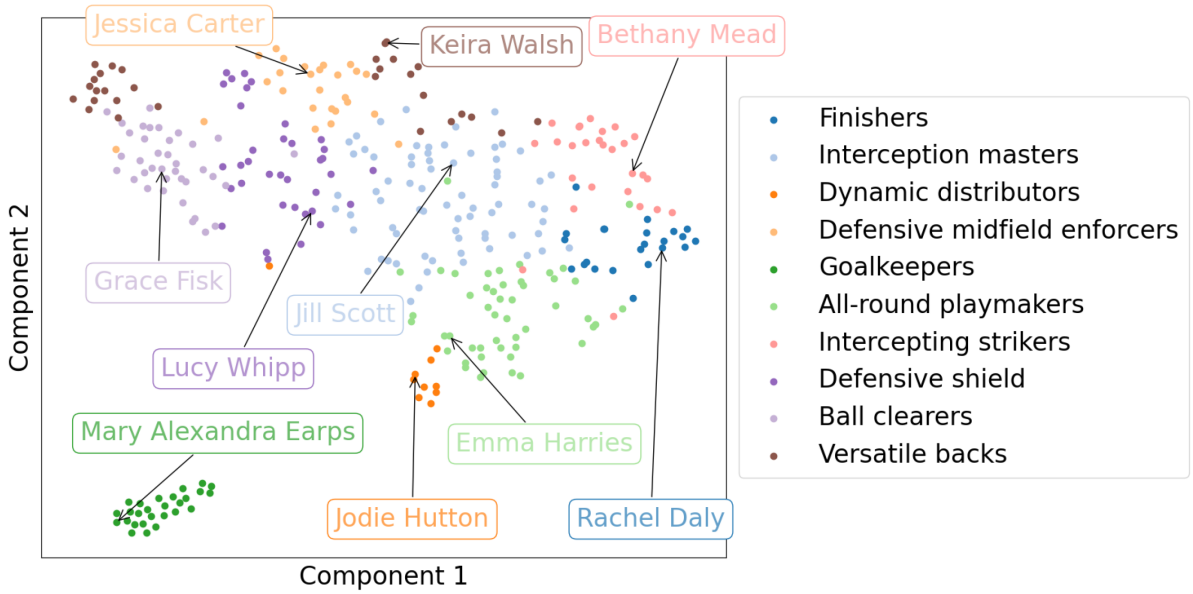



Figure 1: Clusters visualed over the two first components of the PCA in [6]

## 3.1 Algorithms

**K-means** K-means clustering is a popular unsupervised learning algorithm used for partitioning a dataset into $K$ clusters. The goal is to minimize the within-cluster variance, where each cluster is represented by its centroid, a point that minimizes the sum of squared distances from all points in the cluster. The algorithm iteratively assigns each data point to the nearest centroid and updates the centroids based on the mean of the points assigned to each cluster. This process continues until convergence, where the assignments and centroids stabilize, or until a predefined number of iterations.

Formally, let $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$ be the dataset with $n$ data points in a $d$-dimensional space. The algorithm seeks to minimize the objective function:

$$J = \sum_{i=1}^{n} \min_{j} ||\boldsymbol{x_i} - \boldsymbol{\mu_j}||^2$$

where $\mu_j$ is the centroid of cluster $j$, and $|| \cdot ||$ denotes the Euclidean distance.

In this work we iteratively choose initial centroids that are farther away from previously selected ones, aiming to improve the convergence and quality of the final clustering, aslo called K-means++ [7].

**Gaussian Mixture Models** Gaussian Mixture Models (GMMs) are probabilistic models used for representing the distribution of data as a mixture of several Gaussian distributions. Each Gaussian component represents a cluster in the data, and the model parameters include

the mean, covariance, and mixing coefficients for each component. Given a dataset $X = \{x_1, x_2, ..., x_n\}$, GMM assumes that each data point $x_i$ is generated from one of $K$ Gaussian distributions with probabilities governed by mixing coefficients $\pi_k$ and parameters $\{\mu_k, \Sigma_k\}$, where $\mu_k$ is the mean and $\Sigma_k$ is the covariance matrix of the $k$-th Gaussian component.

The probability density function of GMM is expressed as:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the density function of a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

GMMs are often estimated using the Expectation-Maximization (EM) algorithm, which iteratively updates the parameters to maximize the likelihood of the observed data. In this work, we will be using the R package `mclust` for implementing GMM clustering [8].

**DBSCAN** DBSCAN [9] is a density-based clustering algorithm that groups together data points that are closely packed, while marking points in low-density regions as outliers or noise. It defines clusters as continuous regions of high density separated by regions of low density. Given a dataset $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$, DBSCAN defines two parameters: $\varepsilon$, the maximum distance between two points to be considered as neighbors, and $minPts$, the minimum number of points required to form a dense region (cluster). DBSCAN operates by iteratively exploring the neighborhood of each data point. Points with a sufficient number of neighbors within distance $\varepsilon$ are considered core points and form the core of a cluster. Points that are reachable from core points but do not have enough neighbors to be core themselves are considered border points and are assigned to the nearest core point's cluster. Points that are not core or border points and do not belong to any cluster are considered noise points. DBSCAN is robust to noise and can identify clusters of arbitrary shapes and sizes without prior knowledge of the number of clusters.

**Autoencoders** Autoencoders are neural network models used for unsupervised learning that aim to learn efficient representations of input data by reconstructing the input from a compressed latent space. An autoencoder consists of an encoder network that maps the input data to a lower-dimensional latent space representation and a decoder network that reconstructs the input data from this representation. The network is trained to minimize the reconstruction error, typically measured as the difference between the input and output data. We usually use the binary cross-entropy loss or the mean squared error as the reconstruction loss. Autoencoders can be applied to clustering by leveraging the latent space representations learned during the training process. Once the autoencoder is trained on the input data, the encoder network can be used to map data points to a lower-dimensional latent space. Clustering algorithms such as K-means or DBSCAN can then be applied to the latent space representations to group similar data points together. This approach enables unsupervised clustering of high-dimensional data by first reducing the dimensionality using autoencoders,

which can capture complex nonlinear relationships in the data. Autoencoders have been shown to be an extension of PCA with non-linear transforms.

### 3.2 Metrics

In our exploration of clustering techniques, we employed several metrics to evaluate the efficacy and quality of the clustering results. For Gaussian Mixture Models (GMM), we utilized the Bayesian Information Criterion (BIC), defined as:

$$\text{BIC} = -2\log(L) + k\log(n)$$

where $L$ is the likelihood of the data given the model, $k$ is the number of parameters in the model, and $n$ is the number of data points. BIC helps in determining the optimal number of clusters by balancing the goodness of fit with the complexity of the model.

For K-means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), we employed the elbow method. The elbow method aids in determining the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters. The point at which the rate of decrease in WCSS slows down, forming an "elbow" shape in the plot, indicates an optimal number of clusters.

In the case of Autoencoder (AE) based clustering, we first assessed the reconstruction error, defined as:

$$\text{Reconstruction Error} = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2$$

where $\boldsymbol{x}_i$ is the input data point, $\hat{\boldsymbol{x}}_i$ is its reconstruction obtained from the autoencoder, and $N$ is the number of data points. Subsequently, for further evaluation of the clustering performance, we utilized clustering validation indices such as the Davies-Bouldin Index (DBI) or the Silhouette score. The DBI is calculated as:

$$\text{DBI} = \frac{1}{K}\sum_{k=1}^{K} \max_{k \neq k'} \left( \frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(m_k, m_{k'})} \right)$$

where $K$ is the number of clusters, $m_k$ is the centroid of cluster $k$ , $\bar{\delta}_k$ is the average distance from $m_k$ to the points of cluster $k$ and $d(m_k, m_{k'})$ the distance between the centroids of clusters $k$ and $k'$. The Silhouette score is given by:

$$\text{Silhouette Score} = \frac{1}{N}\sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}$$

where $a_i$ is the average distance from $x_i$ to other points in the same cluster, and $b_i$ is the smallest average distance from $x_i$ to points in a different cluster.

## 4 Experiments

In this section we present the StatsBomb data used in this study and the results obtained.

## 4.1 Data

StatsBomb [10] offers a comprehensive dataset covering various facets of football analysis, including event data, tracking data, advanced metrics, player and team performances and many more. The event data, in particular, provides a granular breakdown of every on-ball action occurring during matches (more than 3000 per game). These actions include passes, shots, tackles, interceptions, fouls, and more (see Table 1). Each event is accompanied by detailed attributes such as the player involved, the location on the pitch, the outcome, and additional contextual information. This rich dataset empowers researchers and analysts to delve deep into player performance analysis, team tactics, and game strategies. This dataset will be used through various Women's football competitions to characterise different player types. This study focuses on 473 matches that involved women's football players. Only players with more than 1000 minutes played across these matches were kept (198 players).

| **Variable** | **Meaning** |
|---|---|
| Event (e.g, Pass, Carry) | Average number of occurences per 90 minutes played |
| pass_lengthvar | Variance of pass distance (in yards) |
| pass_lengthmean | Mean of pass distance (in yards) |
| pass_anglevar | Variance of pass angle (in rad) |
| pass_anglemean | Mean of pass angle (in rad) |
| durationvar | Variance of carry duration (in seconds) |
| durationmean | Mean of carry duration (in seconds) |
| Ground Pass | Proportion of Ground Passes |
| High Pass | Proportion of High Passes |
| Low Pass | Proportion of Low Passes |

Table 1: Relevant variables used for clustering

## 4.2 Results

It appears that the various clustering methods returned an optimal number of clusters around 21 (see Figure 2). With this number of clusters used in the kmeans algorithms, the feature pass_lengthvar emerges as a key contributor (see Figure 3), as indicated by its high variance ratio between the within-cluster sum of squares to total sum of squares for each feature. This suggests that the variability in passing lengths plays a significant role in distinguishing between clusters. Despite thorough exploration, the DBSCAN and Autoencoder models used in this study did not demonstrate superior performance or offer additional insights beyond those obtained from K-means and Gaussian Mixture Models (GMM).

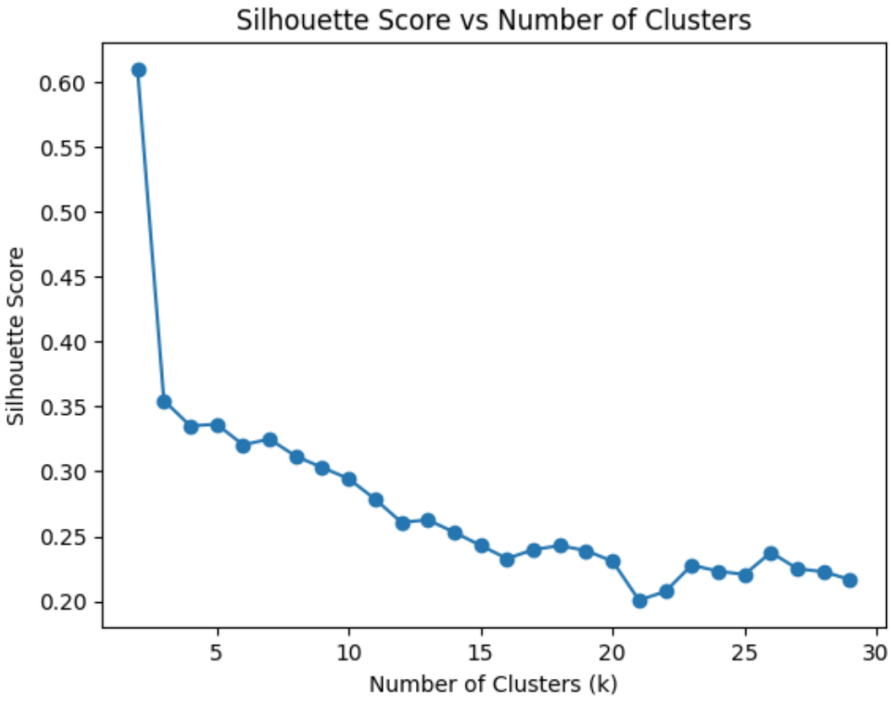



Figure 2: Silhouette scores for $k \in 2 \ldots 30$ clusters with kmeans

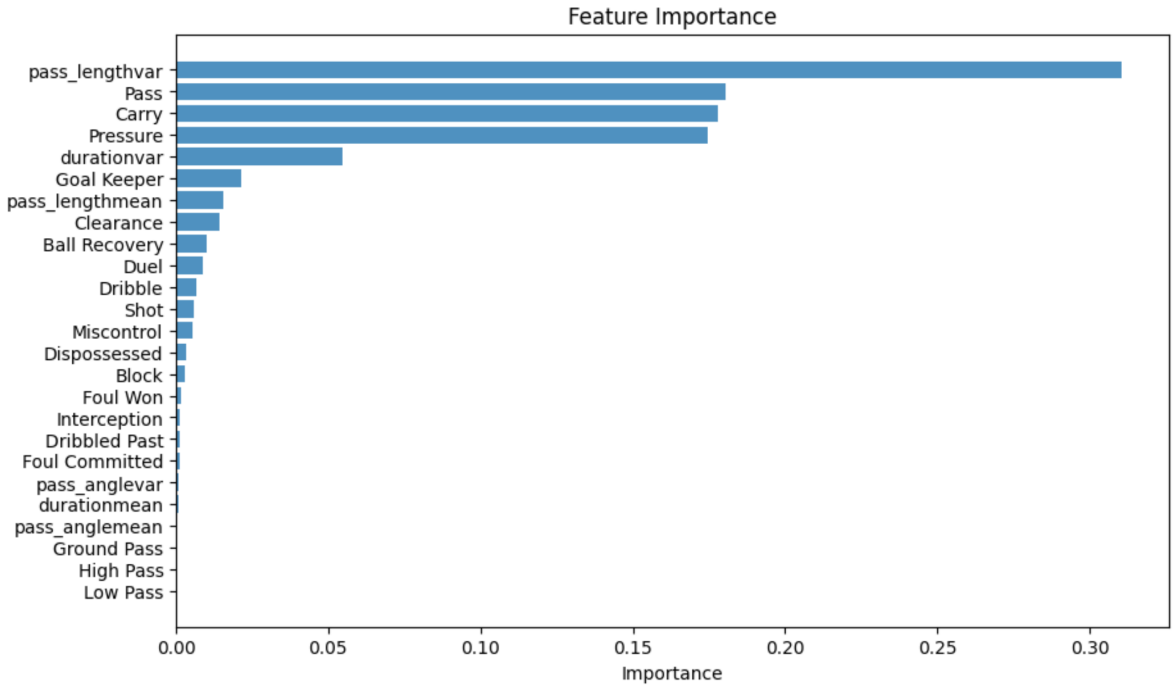



Figure 3: Feature importance calculated using the variance ratio between the within-cluster sum of squares to total sum of squares for each feature

# 5 Conclusion

In conclusion, this study proposed an initial exploration into characterizing women's football players with clustering methods. While our findings provide valuable insights into clustering methods and prominent features such as pass length variance, it is important to recognize the preliminary nature of this work. Future research endeavors could delve deeper into this, focusing on refining clustering techniques and incorporating additional features to enhance player characterization. Notably, our observations suggest that pass length variance serves as a robust indicator of the diversity in players' styles, with strikers demonstrating preference for playing in confined spaces, midfielders exhibiting a mix of short and long passes, and defenders favoring longer passes. Further investigations into these nuances could yield valuable implications for player development and tactical strategies in women's football.

# References


[1] SportsPro, "Women's world cup 2023: Attendance figures, viewership & social media," 2023.

[2] J. Williams and R. Hess, "Women, football and history: international perspectives," *The international journal of the history of sport*, vol. 32, no. 18, pp. 2115–2122, 2015.

[3] M. H. Engh, "Tackling femininity: The heterosexual paradigm and women's soccer in south africa," in *Sport Past and Present in South Africa*, pp. 136–151, Routledge, 2013.

[4] V. Martinez-Lagunas, M. Niessen, and U. Hartmann, "Women's football: Player characteristics and demands of the game," *Journal of Sport and Health Science*, vol. 3, no. 4, pp. 258–272, 2014.

[5] K. Goranovic, A. Lilić, S. Karišik, N. Eler, M. Anelić, and M. Joksimović, "Morphological characteristics, body composition and explosive power in female football professional players.," *Journal of Physical Education & Sport*, vol. 21, no. 1, 2021.

[6] M. Trower, N. Graham, N. Cottrell, and Y. Hengster, "Clustering women's football players," *StatsBomb Conference*, 2023.

[7] D. Arthur, S. Vassilvitskii, *et al.*, "k-means++: The advantages of careful seeding," in *Soda*, vol. 7, pp. 1027–1035, 2007.

[8] C. Fraley, A. E. Raftery, L. Scrucca, and M. L. Berrendero, *mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*, 2022. R package version 5.4.14.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231, 1996.

[10] Benjamin Robinson, *statsbombR: R Client for the 'StatsBomb' API*, 2022. R package version 0.5.1.

# Réduction de dimension

# Gaussian process regression based on dimension reduction for time series outputs

Baptiste Kerleguer [1]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France & baptiste.kerleguer@cea.fr*

**Résumé.** La régression par processus gaussien est largement utilisée pour émuler la sortie d'un code coûteux. Nous nous intéressons à des codes dont la sortie est une série temporelle. Pour réaliser des métamodèles de ces codes, il est usuel de réduire la dimension. Ensuite, une régression est réalisée dans l'espace latent, par exemple régression par processus Gaussien. Ce travail étudie l'influence de l'indépendance des paramètres dans l'espace latent sur les sorties. Pour cela, à l'aide de données simulées et d'un système physique, nous étudions les processus ainsi créés dans l'espace latent et nous les comparons aux hypothèses de décorrélation des coefficients dans l'espace latent faites lors de la métamodélisation.

**Mots-clés.** Reduction de la dimension, Processus Gaussien, Métamodèle

**Abstract.** Gaussian process regression is commonly used to emulate the output of an expensive code. The code outputs are assumed to be a time series. To produce surrogate models of these codes, it is common practice to reduce the dimension. Regression is then performed in the latent space using, for example, Gaussian process regression. This work studies the influence of the independence of the parameters in the latent space. We use simulated data and a physical system. We study the processes created in the latent space and compare them with the assumptions made during surrogate modeling (decorrelation and same variance).

**Keywords.** Dimension reduction, Gaussian process, surrogate model

## 1 Introduction

Advances in scientific modeling have led to the development of more complex and computationally expensive codes. It has therefore become necessary to use surrogate models, constructed from the outputs of these codes, in order to study the uncertainties of these codes. One method widely used in the uncertainty quantification community to produce surrogate models is Gaussian process (GP) regression, see Gramacy (2022) and Williams and Rasmussen (2006). This method, also known as Kriging, was introduced for geostatistics before being used for numerical experiments and uncertainty quantification.

As the complexity of scientific modeling has increased, model outputs have become more and more sophisticated. One of the difficulties is the high dimension of the outputs. Among the codes with high-dimensional outputs, we are interested in those whose outputs are time-dependent functions. When they are sampled, they are called time series. We assume that

the sampling is done on a fine, regular grid, so the output is very high-dimensional. It is assumed that time series are available at all times for each call to the code.

When the regression problem has high-dimensional outputs, such as time series, then the natural solution is to reduce the dimension of the outputs to return to the previous case. The problem is therefore divided into two parts: dimension reduction and Gaussian process regression. Usually, these two parts are treated one after the other, with a certain degree of decoupling. The question is: to what extent is it possible to have a regression model that is independent of dimension reduction? In other words, can we change the dimension reduction method without changing the model in latent space?

The problem we are interested in is the following: let us consider a computational code of which outputs are of the form $z(\mathbf{x}, t)$ with $\mathbf{x} \in \mathbb{R}$ and $t \in [0, 1]$. In the following the stochastic process emulating the code in $Z(\mathbf{x}, t)$. We know this code in a limited number of $\mathbf{x}$ but for $t$ discretized on a regular grid. We try to reduce the output $z(\mathbf{x}, t)$ by $r$ coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$. For that we use a dimension reduction noted $\mathcal{F}$. The principle of dimension reduction is illustrated in Figure 1. For the regression, the objective is to predict in the space of $z(\mathbf{x}, t)$, it is thus necessary to be able to return to the output space of the code. This is how we look for a pseudo-inverse of $\mathcal{F}$ denoted $\tilde{\mathcal{F}^{-1}}$. Note that $\tilde{\mathcal{F}^{-1}}$ is not an inverse of $\mathcal{F}$ because there would be no reduction of dimension. The only exception would be the case where $r$ is equal to the number of points on the time grid. Thus, we obtain an approximation $\hat{z}(\mathbf{x}, t)$ of $z(\mathbf{x}, t)$. Our objective is to have $r$ as small as possible and to have $\hat{z}(\mathbf{x}, t)$ as close as possible to $z(\mathbf{x}, t)$. In order to solve the regression problem, GP regression will be used on the coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$. The space of these coefficients is called the latent space.

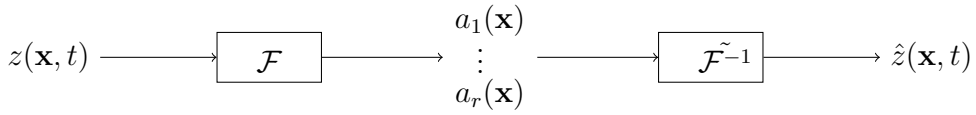



Figure 1: Illustration of the dimension reduction for time series outputs. $\mathcal{F}$ is the dimension reduction function and $\mathcal{F}^{-1}$ its pseudo inverse.

A survey of surrogate model methods for high-dimensional outputs with uncertainty quantification is available in Kontolati et al. (2022). We restrict ourselves to methods that allow an inverse reconstruction in the original space (existence of a pseudo-inverse introduced before).

All that remains, given the dimension reduction, to emulate the coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$ to obtain a surrogate model of the output. We assume that $\mathbf{A} = \{A_1(\mathbf{x}), \cdots, A_r(\mathbf{x})\}$ is a Gaussian process for which the coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$ are a realisation. The common assumption for surrogate modeling the vector $\mathbf{A}$ is that its coefficients are uncorrelated. With an appropriate covariance kernel it is possible to predict $\mathbf{A}$ and therefore to have a surrogate model of $Z(\mathbf{x}, t)$.

In the following section, we present the pendulum attached to a mass-spring system we will use to illustrate this work. Then, in the third section, we study Gaussian processes in the latent space. Using the dimension reduction part of these methods, we will then see how the stochastic process $Z(\mathbf{x}, t)$ behaves compare to a Gaussian process with known mean and

covariance.

## 2 Physical data

Figure 2 shows the system we propose to study. The data comes from a numerical simulation. For the numerical simulation the spring is assumed to have a restoring force proportional to $k$ and the pendulum is assumed to have a mass $m$ at one point. The inputs are $k$, $M$, $\theta$, $\dot{\theta}$ and $y_0$. The output is the position of the pendulum during the first 10 seconds. The other values are fixed. The system with the variations presented was introduced in Perrin (2020). A study of this system with the concidered ranges of variations was carried out in Kerleguer (2023). Perrin (2020) shows that it is possible to build a efficient surrogate model of this system either by dimension reduction and GP regression or by tensorization of the covariance on a regular grid for GP regression.

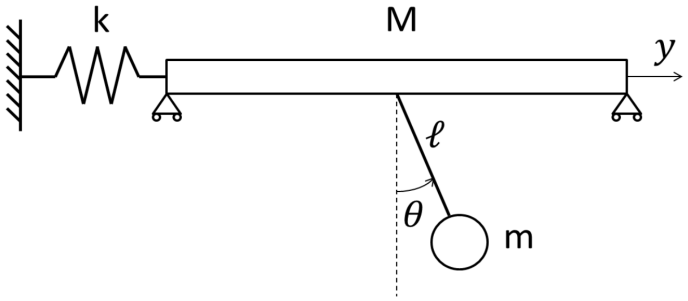



Figure 2: Diagram of a mass-spring system with a simple pendulum attached to it.

This model is used to construct the data studied below. To do this, a sample of 1500 realizations of the system drawn uniformly in the space of the inputs is carried out. The temporal properties of the output signal are given in Figure 3 and 4. Figure 3 shows that the system has a non-zero trend. Furthermore, the convariance, shown in Figure 4, seems to indicate that the covariance is time-dependent.

All the figures of this document showing stochastic processes use the same pattern. Figures 3 and 5 to 9 show stochastic processes with 10 realisations in colour and the mean in dotted grey. The 95% interval centred on the mean is shown in grey.

## 3 Gaussian process in the latent space

In this section, we present dimension reduction methods used in more detail. We then propose a method for understanding the behavior of Gaussian processes in latent space. Then, two approaches to constructing stochastic processes in latent space are compared. The first approach consists of generating Gaussian processes in real space and then passing them into

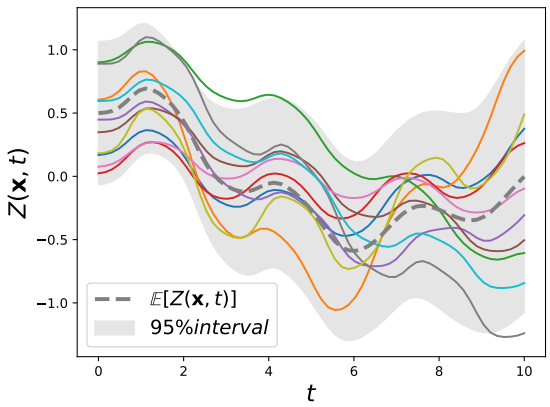



Figure 3: Output of the mass-pendulum system

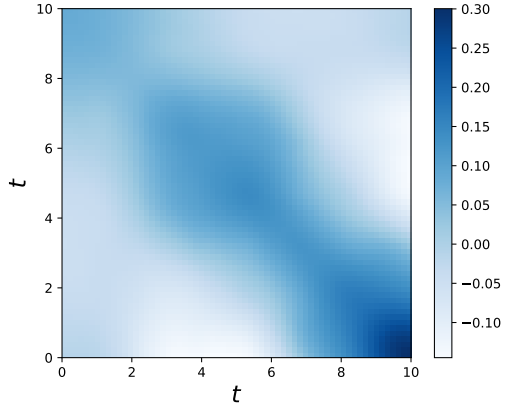



Figure 4: Covariance of the time output of the system

latent space. The process laws are then studied in latent space. The second approach is to assume that the coefficients are independent in the latent space. We can then generate realizations of a Gaussian vector in latent space and study the processes in arrival space. Finally, the GPs are compared with the pendulum data.

## 3.1 Dimension reduction

The different dimension reduction techniques for surrogate modeling used in this work are:

- Singular value decomposition (SVD), with GP regression presented in Nanty (2017),
- Wavelet transform, with GP regression applications given in Perrin et al. (2021) and Rohmer et al. (2023),
- Autoencoder, with an application in Donnelly et al. (2024).

These three types of dimension reduction are studied because they make it easy to find $\tilde{\mathcal{F}^{-1}}$. In addition, these methods are widely used in the literature, see Kontolati et al. (2022). Moreover, studies on the latent space in the case of wavelets with computation of the corresponding covariance kernel have already been carried out in Kerleguer (2022).

The SVD gives an expression of the Gaussian process $Z$ depending on the basis and coefficients:

$$Z(\mathbf{x}, t) = \sum_{i=1}^{r} A_i(\mathbf{x})\Gamma_i(t) + \varepsilon(\mathbf{x}, t), \tag{1}$$

with $N$ is the number of elements, $\mathbf{\Gamma}_N = \{\Gamma_i\}_{i=1,\dots,N}$ is the basis, the coefficients are $A_i$ and $\varepsilon$ is the error of dimension reduction. The base $\mathbf{\Gamma}_N$ is computed using the training data. The

decomposition is then performed each time the surrogate model is called. The equation (1) can be detailed for the case of wavelet decomposition. The equation thus becomes:

$$Z(\mathbf{x}, t) = \sum_{m,n \in \mathcal{I}} A_{i=\{m,n\}}(\mathbf{x}) \psi_{m,n}(t) + \varepsilon(\mathbf{x}, t), \tag{2}$$

with $I$ the space of the wavelet coefficients, and $\psi$ the wavelet function. In the following, the Haar decomposition will be used.

For autoencoder, a $\mathbf{\Gamma}_N$ basis representing the latent space is not available. This complicates the expression of the equations (1) & (2). However, it is always possible to return from latent space to time series space. To compute the autoencoder, the network is small: 2 hidden layers with 20% dropout during training. Optimization time is of the order of a minute on a desktop computer. The latent space of the autoencoder is of dimension 16.

## 3.2 Gaussian process in dimension reduction

$Z(\mathbf{x}, t)$ is assumed to be a Gaussian process with mean zero and covariance Matérn kernel $5/2$ and $\mathbb{V}[Z(\mathbf{x}, t)] = 1$. This process is then applied to the various dimension reduction methods. This gives figures 5, 7 and 9 for SVD, autoencoder and wavelets respectively. The two linear methods show similar behaviour and a rapid decrease in the importance of the coefficients. This decrease is faster for the SVD method, which uses the data to calculate an adapted base. For the autoencoder, the behaviour is different and the coefficients all seem to behave similarly. As the coefficients are exchangeable in the model, this is also to be expected.

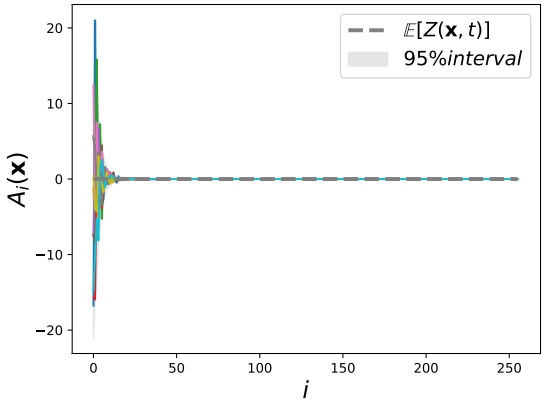



Figure 5: Gaussian Process in the SVD latent space

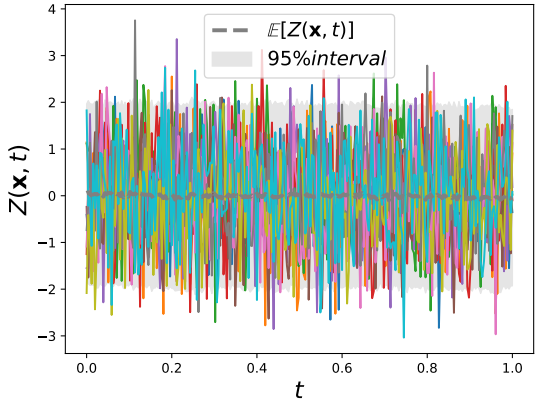



Figure 6: Prediction of a process in the time space with SVD

Figures 6, 8 and 10 are constructed assuming that $A = [A_1(x), \ldots, A_N(x)]$ is a Gaussian vector with mean zero and covariance identity. We can see that for the SVD and wavelet methods, we have high-frequency curves in the time series space. This is because, in the case of a regression, a truncation would be performed, which would eliminate the high-frequency

coefficients. In addition, the decrease in variance will lead to coefficients becoming less and less important as it increases. The mean and variance of the processes are coherent with the input process. In the case of the autoencoder, the samples are much closer to Gaussian processes with Matèrn kernel. However, the variance is lower by an order of two, as can be seen from the 95% interval.

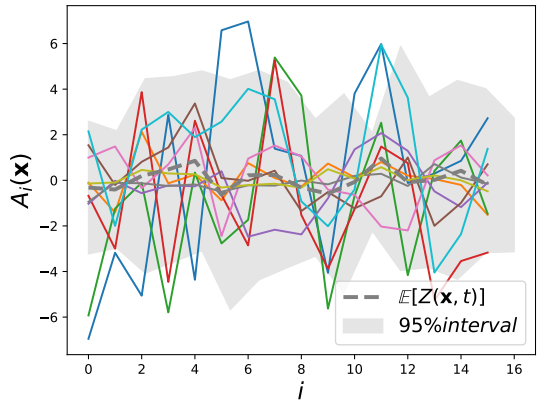



Figure 7: Gaussian Process in the autoencoder latent space

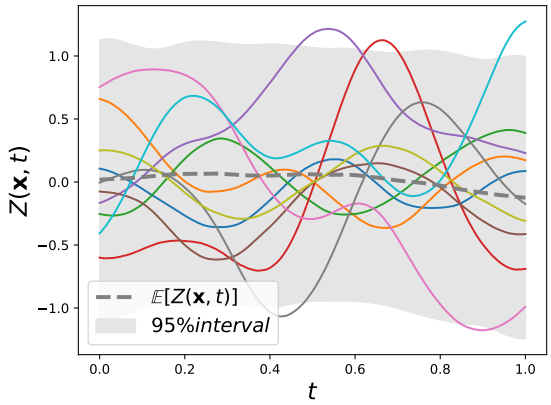



Figure 8: Prediction of a process in the time space with autoencoder

In this example, there is a close relationship between SVD and wavelet decomposition. The latter being a deterministic decomposition, this motivates the approach presented in Perrin (2021).

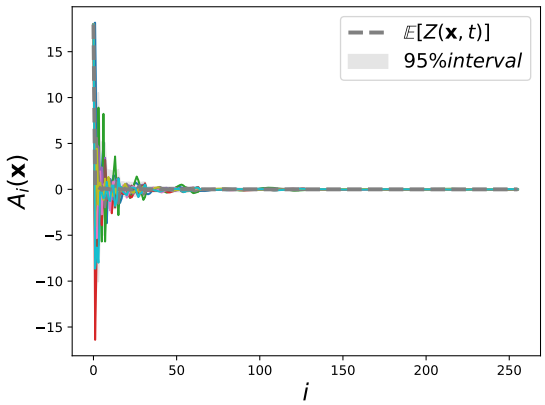



Figure 9: Gaussian Process in the wavelet space

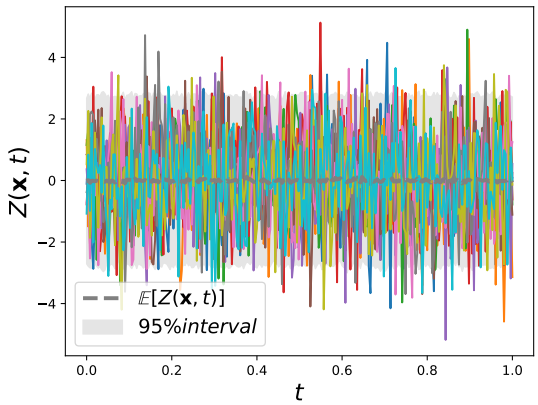



Figure 10: Prediction of a process in the time space with wavelet decomposition

The covariance matrices of the dimension reduction methods also explain the differences between the methods. For SVD, we obtain a diagonal matrix, as expected in theory. The diagonal has decreasing coefficients, which are rapidly negligible numerically. The wavelet

decomposition also shows rapid decrease on the diagonal. Note that this decrease is also visible in Figures 2 and 6. However, while many non-diagonal terms are zero as predicted in Morel et al. (2023), there are still some non-zero terms, see Figure 11, but it seems possible to neglect it. The analytical calculation of these terms is available in Kerleguer (2022). For autoencoding, the covariance matrix is full, see Figure 12, as you would expect from a non-linear method. However, this has no influence on the model's ability to generate Gaussian processes in the time series space, see Figure 5.

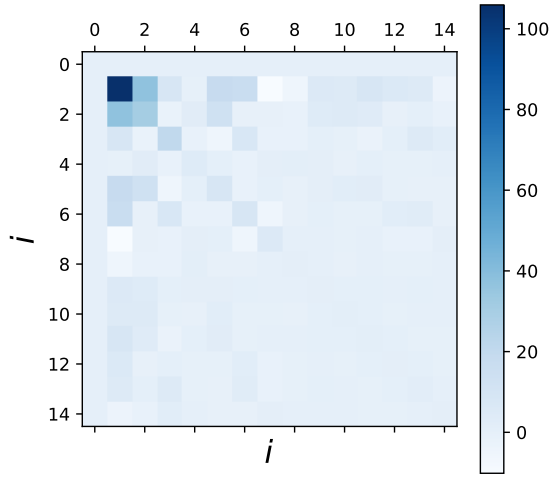



Figure 11: Covariance in the wavelet space

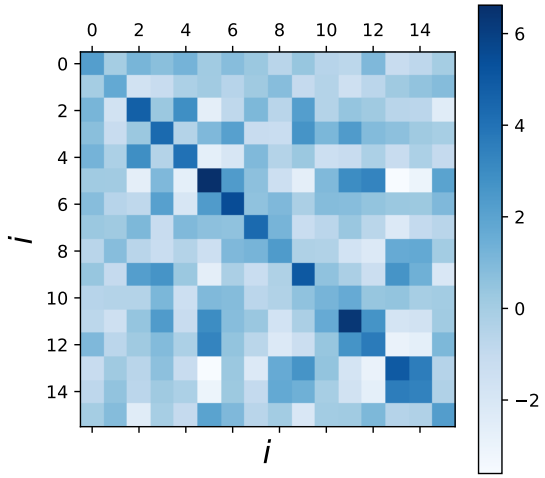



Figure 12: Covariance in the autoencoder space

## 3.3 Application of methods to pendulum systems

Now that the assumptions made in the dimension-reduced GP surrogate modeling have been fixed, these assumptions are compared with the pendulum data. Figures 13, 14 and 15 show the same methods as Figures 5, 7 and 9 but for pendulum data. It should be noted for Figures 13 and 14 that the latent space is of a smaller dimension than the latent space of a GP because the data is less sampled than the realizations of the GP. The mean is the same in the latent space for pendulum data as for GP, but the variance is much smaller. Consequently, it will be straightforward to model $A$ processes in latent space by GP. The surrogate models built using these 3 dimension reduction methods give an $R^2$ scores for 200 training time series between 0.99 and 1, with a test base of 1300 time series.

These results seem to validate the independence hypotheses proposed in Kontolati et al. (2022). However, the empirical covariance matrix in wavelet space proposed in Figure 16 shows a strong correlation between coefficients. The same is true for the autoencoders, although this problem has already been identified for the GP. This is likely to be a problem

in cases of small data sets where the quantification of uncertainties is important. The $A_i$ correlation should be taken into account in the creation of a surrogate model quantifying the uncertainties.

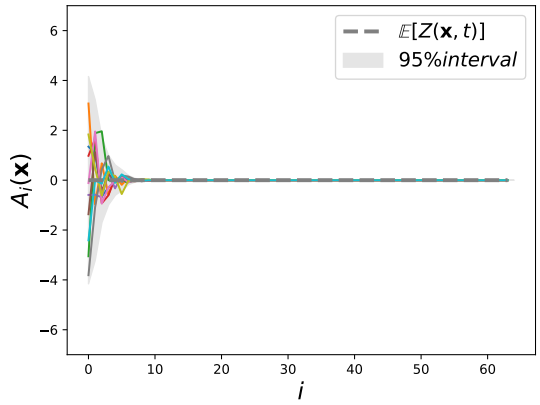



Figure 13: Pendulum data in the SVD

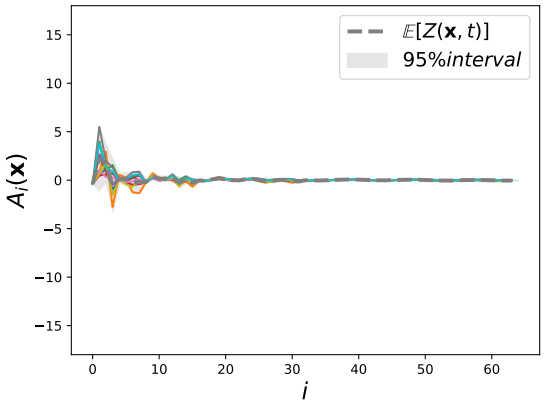



Figure 14: Pendulum data in the wavelet space

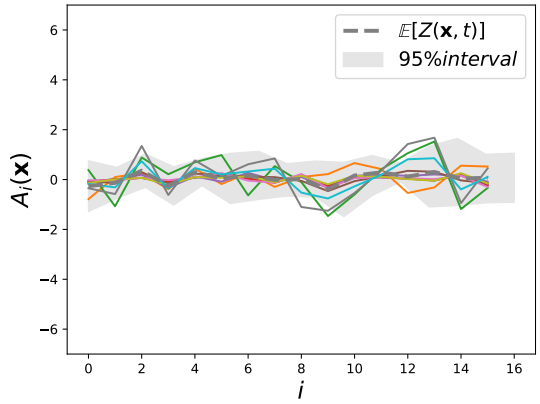



Figure 15: Pendulum data in the autoencoder latent space

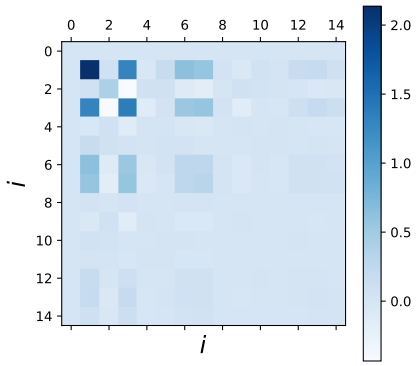



Figure 16: Pendulum data covariance in the wavelet space

# 4 Conclusion

To perform a regression with a time series output, it is possible to reduce the dimension and then perform a Gaussian process regression in the latent space. This method has been shown numerically to be effective on numerous occasions, see Kontolati et al. (2022). However, since dimension reduction does not specifically imply that the coefficients are uncorrelated, it is open to question whether this assumption should be made. We have investigated exactly how this assumption can be made. The data used to study the latent space are of two

kinds: simulated data and data from a physical system. In both cases, it was shown that the processes constructed from these hypotheses can be used for time series regression. However, these hypotheses are not always verified and the quantification of uncertainties in surrogates models could benefit from correlated GP in latent space. In a future work we would like to construct covariance kernels adapted to the different latent spaces in order to be able to take into account correlation between coefficients.

# Bibliographie


Donnelly, J., Daneshkhah, A., & Abolfathi, S. (2024). Forecasting global climate drivers using Gaussian processes and convolutional autoencoders, *Engineering Applications of Artificial Intelligence*, 128, 107536.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*, CRC press, Boca Raton.

Kerleguer, B. (2022). Multi-fidelity surrogate modeling adapted to functional outputs for uncertainty quantification of complex models *(Doctoral dissertation, Institut Polytechnique de Paris)*.

Kerleguer, B. (2023). Multifidelity Surrogate Modeling for Time-Series Outputs, *SIAM/ASA Journal on Uncertainty Quantification*, 11(2), 514-539.

Kontolati, K., Loukrezis, D., Giovanis, D. G., Vandanapu, L., & Shields, M. D. (2022). A survey of unsupervised learning methods for high-dimensional uncertainty quantification in black-box-type problems, *Journal of Computational Physics*, 464, 111313.

Morel, R., Rochette, G., Leonarduzzi, R., Bouchaud, J. P., & Mallat, S. (2023). Scale dependencies and self-similar models with wavelet scattering spectra, *Available at SSRN* 4516767.

Nanty, S., Helbert, C., Marrel, A., Pérot, N., & Prieur, C. (2017). Uncertainty quantification for functional dependent random variables,*Computational Statistics*, 32, 559-583.

Perrin, T. V. E., Roustant, O., Rohmer, J., Alata, O., Naulin, J. P., Idier, D., Pedreros, R., Moncoulon, D. & Tinard, P. (2021). Functional principal component analysis for global sensitivity analysis of model with spatial output. *Reliability Engineering & System Safety*, 211, 107522.

Perrin, G. (2020). Adaptive calibration of a computer code with time-series output, *Reliability engineering & system safety*, 196, 106728.

Rohmer, J., Sire, C., Lecacheux, S., Idier, D., & Pedreros, R. (2023). Improved metamodels for predicting high-dimensional outputs by accounting for the dependence structure of the latent variables: application to marine flooding, *Stochastic Environmental Research and Risk Assessment*, 1-23.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, Cambridge, MA: MIT press.

# Analysing discrete and continuous spectrum and dimension reduction for thermal fields

Mélanie Dreina[1], Sylvie Viguier-Pla[1,2] & Stéphane Abide[3]

[1] *LAMPS, Université de Perpignan via Domitia, F-66860 Perpignan Cedex 9, France, melanie.dreina@univ-perp.fr*
[2] *IMT, Université Paul Sabatier, F-31062 Toulouse Cedex 4, France, viguier@univ-perp.fr*
[3] *LJAD, Université Côte d'Azur, CNRS, F-06108 Nice Cedex 2, France, stephane.abide@univ-cotedazur.fr*

**Résumé.** Le but de ce travail est de comparer trois méthodes de réduction des données dans un contexte de transfert de chaleur. Nous nous plaçons dans le cas bien connu où nous observons des phénomènes instables dans le temps et l'espace. Plus précisément, nous nous intéressons à l'analyse en composantes principales (appelée dans le domaine de la mécanique des fluides la décomposition orthogonale aux valeurs propres ou POD), à la POD spectrale (SPOD), et à l'analyse en composantes principales dans le domaine des fréquences. Les méthodes POD et SPOD ont été proposées dans un contexte de mécanique des fluides, alors que la FPCA est nouvellement appliquée à ce domaine. Ainsi, dans ce travail, nous proposons une discussion sur la capacité de la méthode FPCA à se positionner dans une analyse physique multi-échelle.

**Mots-clés.** Simulation numérique directe, analyse en composantes principales, séries chronologiques, stationnarité, mesure aléatoire, analyse spectrale, champ thermique, ...

**Abstract.** The aim of this work is to compare three methods of data reduction in the context of heat transfer. This follows the well-known practice of observing unsteady phenomena according to space or time energetic arguments. Especially, the present study focuses on the efficiency of the Proper Orthogonal Decomposition (POD), Spectral Proper Orthogonal Decomposition (SPOD) and Principal Components Analysis in the Frequency domain (FPCA). In several previous works, both POD and SPOD have been proposed in the context of fluid mechanics while FPCA is been newly applied to this domain. Thus, in this work we provide a discussion on the contribution of the FPCA to deal with multiscale physics.

**Keywords.** Direct numerical simulation, Principal components analysis, Time series, Stationarity, Spectral analysis, Thermal field,...

## 1 Introduction

Simulation of complex systems, such as in fluid mechanics, leads to the production of a large amount of information. Therefore, dimension reduction is of major importance to be

able to carry out fine analyses of the underlying physical phenomena. In this context, several approaches have been developed since the last decades. One can mention the pioneer works concerning Proper Orthogonal Decomposition(POD) by Lumley (1970), based on the Principal Components Analysis (PCA), or more recently the Dynamic Mode Decomposition (DMD) proposed by Schmid (2010). All of them look for an efficient way to give an alternative representation of data in order to facilitate analysis. Keeping in mind this purpose, PCA is devoted to the reduction of dimension, mainly in a context of independent observations. When data are time-dependent measurements, the independence is no more ensured in the time domain, but when the signal is a stationary process, its Fourier Transform gives independent observations. The Spectral POD (SPOD) is a new method introduced by Lumley (2007) in order to rank the modes according to their energy level having a characteristic frequency. Finally, PCA in the frequency domain (Brillinger (2001)) aims a similar purpose to the SPOD, with PCA of the Fourier Transform for each identified frequency. Boudou (1995) has proposed a generalisation of this PCA in the frequency domain, that we name FPCA, and which has first been performed in Boudou et al. (2004) for periodic flows.

In this study, we propose to compare results from POD, SPOD and an improved method of FPCA, which is not restrictive on the structure of the multidimensional signal spectrum. In this presentation, we first present the data of interest. Secondly, we present the three compared methods, that is POD, SPOD and FPCA. In the third part, we apply the three methods on data from simulation. We end by showing the difficulties of each method, we compare the qualities of reconstruction and the phenomenon each method reveals at each step.

# 2 Description of the interest data

The comparison of POD, SPOD and FPCA methods is carried out on the spatio-temporal serie basis of a natural convection flow temperature field (Sergent et al. (2013), Trias et al. (2007)). In particular, we simulate the thermal coupling between a fluid and a solid wall, imposing continuity of the temperature field at the fluid/solid interface. The data considered here is therefore a sampling of a variable $T(t, x, y)$ determined by Direct Numerical Simulation. For the sake of illustration, a snapshot and a time series are given in Fig. 1. Details of the DNS solver used in this work are presented in Abide (2017, 2018).

The direct numerical simulation of turbulent natural convection at $Ra = 10^9$ (Trias et al. (2007)) has been carried out on a grid $48 \times 65$ over 1000 time steps. In the following, comparisons between POD, SPOD and FPCA are based on a sub-sampling of the wall temperature by 100 snapshots of dimension $24 \times 22$ points.

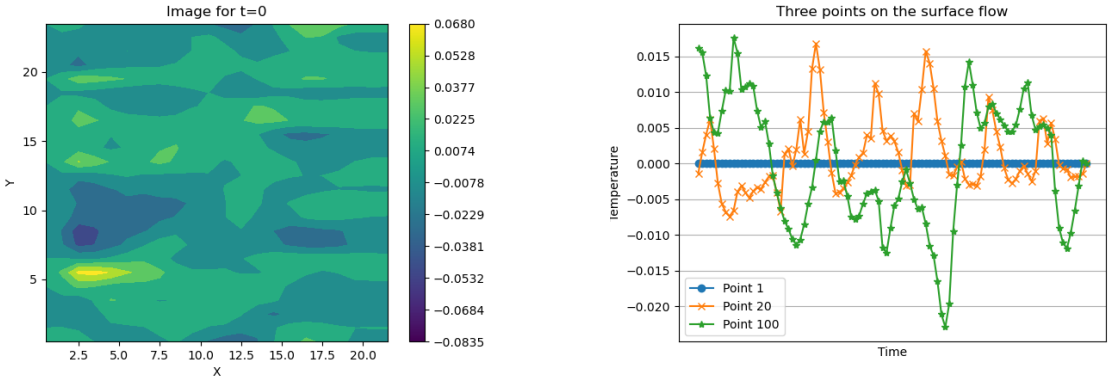



Figure 1: First snapshot and time series of the temperature fluctuations

.

# 3 The methods of dimension reduction

## 3.1 Proper Orthogonal Decomposition

Let $\{u(x,t)\}$ be a stochastic process defined on $\mathbb{R}^n \times \mathbb{R}$, as for example a random $n-$dimensional field observed along the time. The POD is the search of a deterministic function $\phi(x)$ that best approximates the stochastic function in average. Practically, it consists of considering a sample $(x_1, \ldots, x_n)$ of space points, and measures at times $t_1, \ldots, t_p$. The method is implemented via the principal components analysis (PCA) of the matrix $U = (u(x_j, t_i))_{j=1,\ldots,n;i=1,\ldots,p}$. Each principal component of such a PCA is named a mode.

## 3.2 Spectral Proper Orthogonal Decomposition

Spectral proper orthogonal decomposition (SPOD) is a frequency domain variation of POD recently brought to the fore by Towne et al. (2018). This method is designed for statistically stationary flows. It is aimed to extract coherent structures from flow data. For example, it has been applied to extract the spatio-temporal modes of a jet and wind turbine flows in He et al. (2021). The main contribution of the SPOD compared to the POD is that the modes vary in both spatial and temporal dimensions, and are orthogonal under a space-time inner product, as opposed to being purely spatial. Consequently, these modes are optimal for representing spatio-temporal coherence within the data. In mathematical terms, the SPOD modes represent eigenvectors of the cross-spectral density (CSD) matrix at individual frequencies, where the eigenvalues denote the energy associated with each mode at a given frequency. A more detailed description of the method is given by Schmidt & Colonius (2020). We perform the SPOD with the open access python script proposed and successfully applied to several example by He et al. (2021) [https://github.com/HexFluid/spod_python (accessed February 2024)]. This script is built up as follows. The first step is to build a matrix for the spatio-temporal data. Let the vector $q_k \in \mathbb{R}^{N_q}$ be the $k^{th}$ time snapshot after subtracting

the time-averaged data. The chronologically sorted spatio-temporal data matrix is:

$$\mathcal{Q} = [q_1, q_2, \ldots, q_{N_t}] \in \mathbb{R}^{N_q \times N_t},$$

where $N_t$ is the number of snapshots. Secondly, the data matrix is decomposed into $N_b$ blocks using the Welch periodogram method and the discrete Fourier transform is applied to each block to pass into the frequency domain. At this stage, to prevent loss of precision due to spectral leakage, each data block is processed with a Hamming window and then overlapped with neighbouring blocks. The matrix for the $j^{th}$ block is

$$\hat{Q}^{(j)} = \left[\hat{q}_k^{(j)}, \hat{q}_2^{(j)}, \ldots, \hat{q}_{N_f}^{(j)}\right] \in \mathbb{C}^{N_q \times N_f}$$

Then, according to the frequency, the matrices are reshaped so that the matrix for the $k^{th}$ frequency is

$$\hat{Q}_k = \left[\hat{q}_k^{(1)}, \hat{q}_k^{(2)}, \ldots, \hat{q}_k^{(N_b)}\right] \in \mathbb{C}^{N_q \times N_b}.$$

The weighted cross-spectral density (CSD) matrix for the $k^{th}$ frequency, denoted as $S_k$ is obtained as follows:

$$S_k = \frac{1}{N_b} W^{1/2} \hat{Q}_k^* \hat{Q}_k W^{1/2} \in \mathbb{C}^{N_b \times N_b},$$

where $W$ represents the weight matrix for scaling the various flow variables. The specific definition of $W$ determines the physical interpretation of the energy associated with the SPOD modes. Finally, the eigen-decomposition is performed on the weighted cross-spectral density (CSD) matrix $S_k$ for each frequency. The resulting modes are used for a variety of purposes such as classification and reduced order modelling. Similar to other versions of Proper Orthogonal Decomposition (POD), SPOD determines an orthogonal basis for the data, meaning that a subset of these modes captures a proportion of the total energy (variance) within the data compared to any other orthogonal basis. The function used for the reconstruction is based on Nekkanti & Schmidt (2021), and is available in the python script above mentioned.

## 3.3 Principal Components Analysis in the Frequency domain

Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary $p-$dimensional random time series. The FPCA of $(X_n)_{n\in\mathbb{Z}}$ is the search of a $q-$dimensional series $(q < p)$ $(X'_n)_{n\in\mathbb{Z}}$, stationarily correlated with $(X_n)_{n\in\mathbb{Z}}$, as close as possible to it. As $(X_n)_{n\in\mathbb{Z}}$ and $(X'_n)_{n\in\mathbb{Z}}$ are stationary, there exist two unitary operators $U$ and $U'$ such that $X_n = U^n X_0$ et $X'_n = U'^n X'_0$. So the FPCA is the search of $X'_0$ and $U'$ such that $X'_n = U'^n X'_0$ and $\|X_0 - X'_0\|$ is as small as possible.

The $X_n$'s of the stationary series $(X_n)_{n\in\mathbb{Z}}$, are $p-$dimensional random vectors: $X_n = (x_n^1, \ldots, x_n^p)^t$. The stationarity is assumed in a broad sense, that is $\mathbb{E}(X_n\, {}^t\overline{X_m}) = \mathbb{E}(X_{n-m}\, {}^t\overline{X_0})$ for any pair $(n, m)$ of elements from $\mathbb{Z}$. It is equivalent with the usual second order stationarity of each of its components $(x_n^i)_{n\in\mathbb{Z}}$ and with the pairwise correlated stationarity: $\mathbb{E}(x_n^i \overline{x}_m^j) = \mathbb{E}(x_{n-m}^i \overline{x}_0^j)$ for any $(n, m, i, j)$ from $\mathbb{Z} \times \mathbb{Z} \times \{1, ..., p\} \times \{1, ..., p\}$.

We assume that the conditions are satisfied for the existence of the spectral density,

$$(2\,\pi)^{-1} \sum_{n\in\mathbb{Z}} e^{-i.n} \mathbb{E}\, X_n^t \overline{X_0}.$$

Theoretically, the FPCA needs to process the PCA of $(2\,\pi)^{-1}\sum_{n\in\mathbb{Z}} e^{-i\lambda n}\mathbb{E}\,X_n^t\overline{X_0}$, for each $\lambda$ from $[-\pi,\,\pi[$, this means an infinity of PCA's. We overcome this difficulty by a discretization of the spectrum $[-\pi,\,\pi[$. More precisely, if $k$ is an integer, we consider the measurable application from $[-\pi,\,\pi[$ into itself:

$$f_k = \sum_{l=-k}^{k-1} \frac{\pi\,l}{k}\,1_{B_{lk}}$$

where $B_{-k,k} = \{-\pi\}$, $B_{lk} = ]\frac{\pi\,l}{k} - \frac{\pi}{k}, \frac{\pi\,l}{k}]$ for $l = -k+1,\ldots,-1$, $B_{0k} = ]-\frac{\pi}{k}, \frac{\pi}{k}[$, and $B_{lk} = [\frac{\pi\,l}{k}, \frac{\pi\,l}{k}+\frac{\pi}{k}[$ for $l = 1,\ldots,k-1$. The FPCA can be approximated by the spectral decomposition of the spectral density $M_{lk}$ defined on each $B_{lk}$; $l = -k+1,\ldots,k-1$. The matrices $M_{lk}$ can be estimated by

$$(2\,\pi\,m)^{-1}\sum_{u=1}^{m}\sum_{v=1}^{m}(\int_{B_{lk}} e^{i\lambda(u-v)}\,d(\lambda))X_v\,{}^t\overline{X_u}$$

.

Let $(X'_n)_{n\in\mathbb{Z}}$ be the $q-$dimensional solution of the $q-$order FPCA of $(X_n)_{n\in\mathbb{Z}}$. This series is of the form $X'_n = \sum_{m\in\mathbb{Z}} C'_m X_{n-m}$. It can be approximated via the discretization of the spectrum, by the series

$$X'^k_n = \sum_{m\in\mathbb{Z}} C'_{m,k} X_{n-m},$$

where

$$C'_{m,k} = (2\pi)^{-1}\sum_{l=-k+1}^{k-1}(\int_{B_{lk}} e^{i\lambda m}\,d\,\lambda)\sum_{j=1}^{q} F_j\,{}^t\overline{A_{jlk}},$$

$F_j$ being the $j^{\text{th}}$ vector of the canonical basis of $\mathbb{C}^q$, and $A_{jlk}$ being the $j^{\text{th}}$ unitary eigenvector of $M_{lk}$.

The reconstructed series is then $(X''^k_n)_{n\in\mathbb{Z}}$, which can be writen

$$X''^k_n = \sum_{m\in\mathbb{Z}} C''_{m,k} X'^k_{n-m} = \sum_{m\in\mathbb{Z}} D_{m,k} X_{n-m}\ ,$$

where

$$C''_{m,k} = (2\pi)^{-1}\sum_{l=-k+1}^{k-1}(\int_{B_{lk}} e^{i\lambda m}\,d\lambda)\sum_{j=1}^{q} A_{jlk}\,{}^t\overline{F_j}\,,$$

and

$$D_{m,k} = (2\pi)^{-1}\sum_{l=-k+1}^{k-1}(\int_{B_{lk}} e^{i\lambda m}\,d\lambda)\sum_{j=1}^{q} A_{jlk}\,{}^t\overline{A_{jlk}}\,.$$

Of course, the greater is $k$, the nearest the approximated FPCA is to the theoretical FPCA defined above.

We can examine the norms of the $C'_{m,k}$, which are high when the gap $m$ in the linear combination of the reconstruction is high, what happens, for example, when the series is periodic of period $m$. We can also compare the series before and after the FPCA, for various dimension $q$ values of the reconstruction.

# 4 Results and discussion

## 4.1 Analysis with POD

We examine the modes of this analysis, which match with the principal components in usual PCA. In Figure 2, the reconstruction is very slightly improved from 1 to 2 dimensions. At least, the essential of the variations is returned.

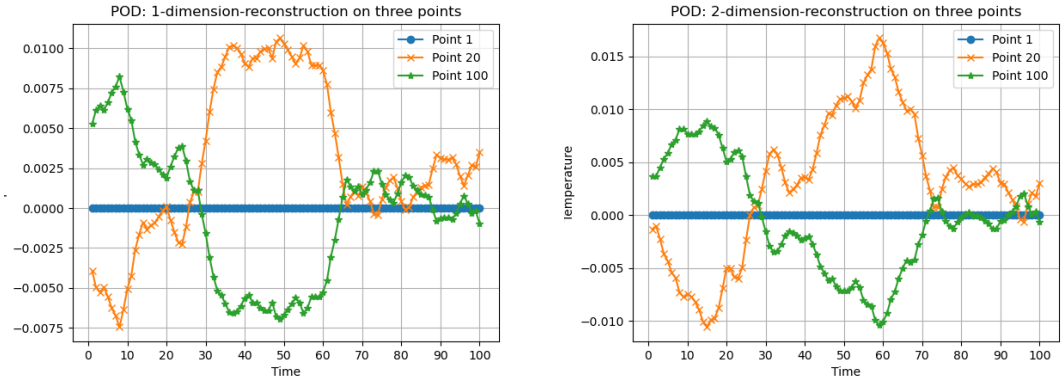



Figure 2: POD : Reconstruction of three points variations with one and two modes

.

The variations most reconstructed are those for median temperature, as we see in Figure 3.

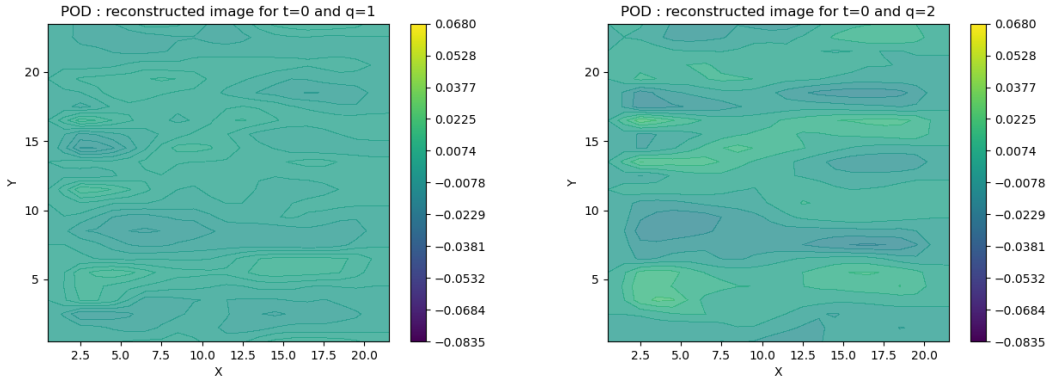



Figure 3: POD : Reconstruction of image at t=0 with one and two modes

.

## 4.2 Analysis with SPOD

In Figure 4, we can see that the variations of points 20 and 100 are slightly more complex than the ones from POD, but the same variations are first retrieved.

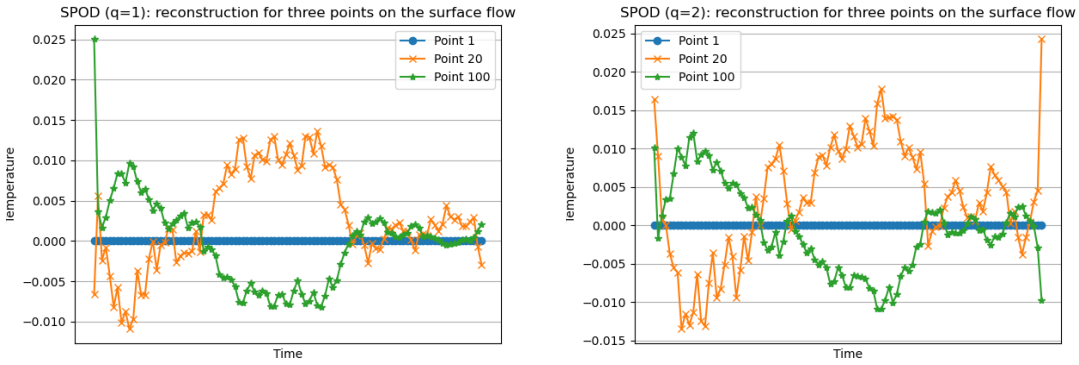



Figure 4: SPOD : Reconstruction of three points variations with one and two modes

.

The variations most reconstructed are those for extreme temperatures at $t = 0$, as we see in Figure 5.

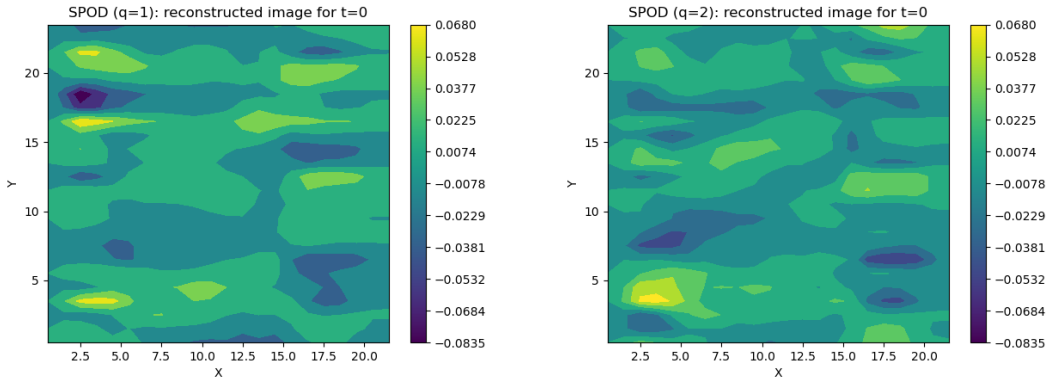



Figure 5: SPOD : Reconstruction of image at t=0 with one and two modes

.

## 4.3 Analysis with FPCA

As for the trajectories of points 20 and 100, FPCA retrieves more complexity than the previous methods. It takes into account more frequencies in the first modes (Figure 6).

Figure 7 gives the first snapshot for a one and two-dimensions reconstruction. By comparison with the initial first snapshot, we can recognise the main variations of the flow, yet for $q = 1$.

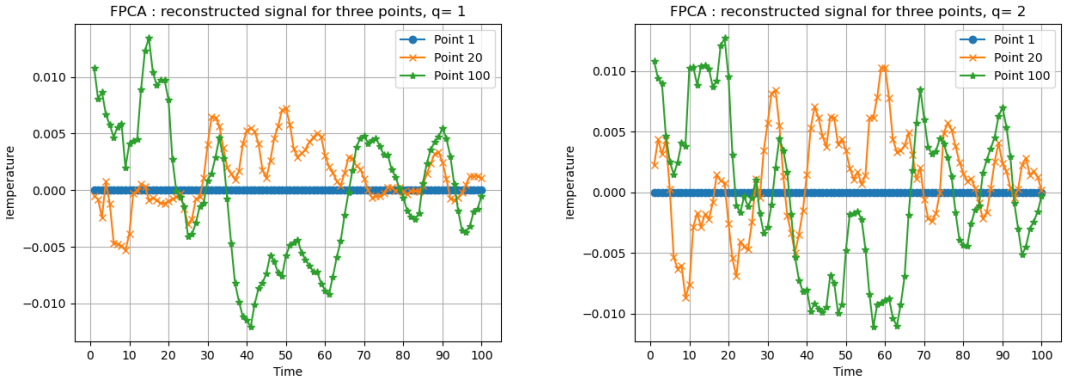



Figure 6: FPCA, $k = 10$ : reconstruction on dimensions 1 and 2

.

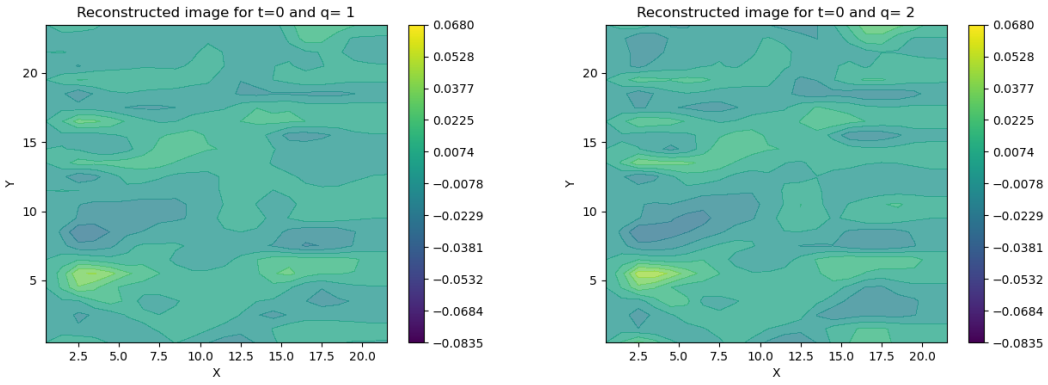



Figure 7: FPCA, $k = 10$ : reconstruction on dimensions 1 and 2

.

## 4.4 Comparisons

One way to assess the efficiency of decomposition method relies on its ability to reconstruct the initial signal with few modes. To this end, we evaluate the error in reconstruction with respect to the mode numbers. Figure 8 presents the relative error computed for the three methods POD, SPOD and FPCA. One can note that FPCA is able to reconstruct the data with fewer modes than the other methods. In this way, FPCA overcomes the POD and SPOD in its ability to retrieve data. Moreover, the control parameter $k$ improves greatly the decomposition efficiency. When $k$ is small, the number of subdivisions of the frequency spectrum is small, so few frequencies are taken into account. The higher $k$ is, the higher is the number of considered potential frequencies. The method FPCA has been performed with $k = 10$ and $k = 20$. This comparison of two values of $k$ illustrates the fact that the higher $k$ is, the smaller the error is, for a fixed value of $q$. As POD and SPOD present similar errors in the first dimensions, SPOD tends to be better with dimension getting higher. FPCA has more little errors, and the quality of reconstruction is almost perfect as soon as the dimension

reaches $q = 10$ when $k = 10$, and $q = 6$ when $k = 20$.

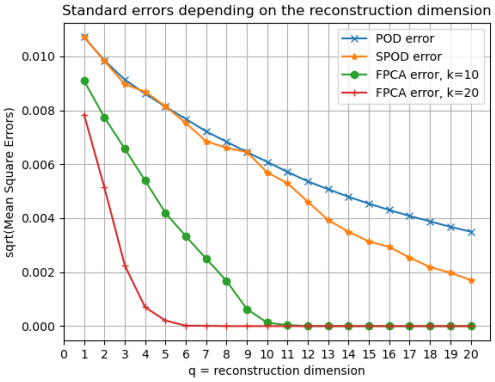



Figure 8: Standard deviation of errors of reconstruction for dimensions 1 to 20

.

# 5 Conclusion

The FPCA sounds interesting for several purposes in fluid mechanics. The summary needs few modes to give good quality of reconstruction compared to POD and SPOD. We can analyze the coefficients of the reconstruction for information above the periodic parts of the signal, and we can select part of the spectrum part for the extraction of some particular phenomena. Moreover, Boudou and Viguier-Pla (2006) have investigated the conditions where PCA and FPCA give the same results. This condition is the independence of data from time, and a consequence of this independence is that FPCA and POD become equivalent. The difference between POD, SPOD and FPCA results give indications about how time-dependent are the data.

FPCA is compared for the first time to SPOD, which is supposed to proceed with the same way of dealing with the frequency domain, and this on data simulated from fluid mechanics models. However, we must also analyze the computational efficiency of each method, and the ability of these methods to apply to large volumes of data. As FPCA has got longer execution time, one of the challenges is to adapt its algorithms to this context.

# Bibliographie

Abide, S. , Binous, M.S. and Zeghmati, B. (2017), An efficient parallel high-order compact scheme for the 3D incompressible Navier-Stokes equations, *International Journal of Computational Fluid Dynamics*, 31, pp. 214-229.

Abide, S., Viazzo, S. and Raspo, I. (2018), Higher-order compact scheme for high-performance computing of stratified rotating flows, *Computers & Fluids*, 174, pp. 300-310.

Boudou, A. (1995), Mise en œuvre de l'analyse en composantes principales d'une série stationnaire multidimensionnelle, *Publications de l'Institut de Statistique de l'Université de Paris*, XXXIX(1), pp. 89-104.

Boudou, A., Caumont, O. and Viguier-Pla, S. (2004), Principal components analysis in the frequency domain, *COMPSTAT 2004-Proceedings in Computational Statistics*, pp. 729-736.

Boudou, A., and Viguier-Pla, S. (2006), On proximity between P.C.A. in the frequency domain and usual P.C.A., *Statistics*, 40, pp. 447-464.

Brillinger, D. R. (2001), *Time Series: Data Analysis and Theory*, Society for Industrial Applied Mathematics, Philadelphia.

He, X., Fang, Z., Rigas, G., and Vahdati, M. (2021), Spectral Proper Orthogonal Decomposition of Compressor Tip Leakage Flow, *Physics of Fluids*, 33(10).

Lumley, J.L. (2007), *Stochastic Tools in Turbulence*, Courrier Corporation, Initially published in 1970 in Academic Press, New-York.

Nekkanti, A., and Schmidt, O.T. (2021), Frequency–Time Analysis, Low-Rank Reconstruction and Denoising of Turbulent Flows Using SPOD, *Journal of Fluid Mechanics*, 926, A26.

Schmid, P.J. (2010), Dynamic mode decomposition of numerical and experimental data, *Journal of Fluid Mechanics*, 656, pp. 5-28.

Schmidt, O.T. and Colonius, T. (2020), Guide to Spectral Proper Orthogonal Decomposition, *AIAA Journal*, 58, pp. 1023–33.

Sergent, A., Xin, S., Joubert, P., Le Quéré, P., Salat, J. and Penot, F. (2013), Resolving the stratification discrepancy of turbulent natural convection in differentially heated air-filled cavities, *International Journal of Heat and Fluid Flow*, 39, pp. 1-14.

Towne, A., Schmidt, O.T. and Colonius, T. (2018), Spectral Proper Orthogonal Decomposition and Its Relationship to Dynamic Mode Decomposition and Resolvent Analysis, *Journal of Fluid Mechanics*, 847, pp. 821–67.

Trias, F. X., Soria, M., Oliva, A. and Pérez-Segarra, C. D. (2007), Direct numerical simulations of two and three-dimensional turbulent natural convection flows in a differentially heated cavity of aspect ratio 4, *Journal of Fluid Mechanics*, 586, pp. 259-293.

# A non asymptotic analysis of the first component PLS regression

Luca Castelli [1] & Irène Gannaz [2]& Clément Marteau[1]

[1] *Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France, castelli@math.univ-lyon1.fr marteau@math.univ-lyon1.fr*

[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France, irene.gannaz@grenoble-inp.fr*

**Résumé.** La régression PLS (*Partial Least Squares*) est une méthode statistique permettant de travailler dans un cadre de grande dimension. Cette méthode projette les covariables sur un sous-espace bien choisi, considérant les corrélations successives avec la variable à expliquer dans le but d'améliorer la qualité de prédiction. Nous nous focaliserons sur le cas de la projection sur une composante, qui fournit un cadre utile pour comprendre le mécanisme sous-jacent. Malgré sa simplicité apparente, ce cadre présente de nombreux défis statistiques. En particulier, la non-linéarité de l'estimateur correspondant exige une attention particulière.

Nous fournissons une borne non asymptotique sur la perte quadratique en prédiction avec grande probabilité. Nous montrons que la qualité de l'estimation PLS dépend de l'inverse de l'inertie relative de la composante par rapport à la variance des covariables. Nous étendons ces résultats à l'approche Sparse PLS. En particulier, nous présentons des bornes supérieures similaires à celles obtenues avec l'algorithme LASSO, avec une contrainte supplémentaire sur les valeurs propres restreintes de la matrice de design.

**Mots-clés.** Réduction de dimension, Régression, Parcimonie

**Abstract.** Partial Least Squares (PLS) regression is a dimension reduction technique used to handle high dimensionality. This method projects the data onto a carefully chosen subspace, considering successive correlations with the explanatory variable in order to improve the prediction quality. We focus our attention on the single component case, that provides a useful framework to understand the underlying mechanism. Despite its apparent simplicity, this scenario presents numerous statistical challenges. Specifically, the non-linearity of the corresponding estimator demands careful attention. We provide a non-asymptotic upper bound on the quadratic loss in prediction with high probability in a high dimensional regression context. The bound is attained thanks to a preliminary test on the first PLS component. In a second time, we extend these results to the sparse partial least squares approach. In particular, we exhibit upper bounds similar to those obtained with the lasso algorithm, up to an additional restricted eigenvalue constraint on the design matrix.

**Keywords.** Dimension reduction, regression, sparsity

# 1 Introduction

Nous nous intéressons au modèle linéaire classique dans un contexte de grande dimension. Nous observons un échantillon de taille $n$, $(X_i, Y_i)$, $i = 1, \ldots, n$, où les $Y_i \in \mathbb{R}$ sont les variables de sortie et les $X_i \in \mathbb{R}^p$ sont les covariables $p$-dimensionnelles. Nous considérons une relation linéaire au sein de chaque couple $(X_i, Y_i)$, représentée par l'équation :

$$Y = X\beta + \varepsilon, \tag{1}$$

où $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim \mathcal{N}_n\big(0, \tau^2 I_n\big)$, $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p}$ et $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$. La matrice $I_n$ est la matrice identité de taille $n$ et le paramètre $\tau$ caractérise le niveau de bruit. Nous désignons par $\Sigma = X^T X/n$ la matrice de Gram associée à la matrice de design $X$, $\hat{\sigma} = X^T Y/n$ la covariance empirique entre $X$ et la cible $Y$, et $\sigma = \mathbb{E}[\hat{\sigma}] = \Sigma\beta$ son espérance.

Notre objectif est d'étudier les performances en prédiction dde la régression par Moindres Carrés Partiels (Partial Least Squares, PLS). L'idée de la régression PLS est de rechercher un nombre fixe de directions - disons $K \in \{1, \ldots, p\}$ - formées par des combinaisons linéaires des coordonnées de $X$, qui sont fortement corrélées avec la variable cible $Y$ (voir Mateos-Aparicio (2011) pour une introduction complète). Ces $K$ directions sont regroupées dans une matrice de poids $W \in \mathbb{R}^{p \times K}$. Le paramètre $\beta$ est ensuite estimé par une combinaison linéaire appropriée des colonnes de $W$. Plus formellement, l'estimateur PLS satisfait :

$$\hat{\beta}_W = \underset{w \in [W]}{\operatorname{argmin}} \ \|Y - Xw\|^2, \tag{2}$$

où $[W] \subset \mathbb{R}^p$ désigne l'espace engendré par les colonnes de la matrice de poids $W$, et $\|.\|$ est la norme $\ell^2$ sur $\mathbb{R}^n$. L'objectif de la réduction de dimension donnée par $W$ est de réduire le nombre de caractéristiques de $p$ à $K$ tout en conservant autant d'informations que possible. Contrairement à la réduction sur composantes principales où les directions sont construites uniquement en considérant les covariables $X$, la régression PLS construit les poids $W$ de manière itérative, en tenant compte des corrélations successives avec la réponse $Y$, afin d'améliorer la qualité de prédiction.

Nous allons concentrer notre attention dans la suite sur le cas de la projection sur une seule composante, c'est-à-dire $K = 1$.

# 2 La régression PLS

La régression PLS a été principalement développée dans la communauté de la chimiométrie (Martens et Naes, 1992). Cette approche a démontré sa capacité à prédire des modèles de régression avec de nombreuses variables prédictives (Garthwaite, 1994). Elle a été largement utilisée en chimiométrie (Wold, 1995; Wold, Sjöström et Eriksson, 2001), mais aussi dans d'autres domaines tels que les sciences sociales (Sawatsky, Clyde et Meek, 2015) et la biologie (Palermo, Piraino et Zucht, 2009; Yang et al., 2017). Plusieurs extensions ont été proposées au fil des ans, comme par exemple Delaigle et Hall (2012) pour les données fonctionnelles ou Naik et Tsai (2000) pour les modèles à indice unique.

La PLS possède une structure algorithmique due à la construction du sous-espace $[W]$ constitué des poids permettant de sélectionner les variables pour favoriser la prédiction.

**Algorithm 1** PLS Algorithm

Input $\mathbf{X}$,Y and K

$\mathbf{X}_1$=$\mathbf{X}$
**for** k=1,..., K **do**
  $\mathbf{w}_k$=$\mathbf{X}^{(k)^T}Y/\|\mathbf{X}^{(k)^T}Y\|_2$ (loadings computation)
  $\mathbf{t}_k$=$\mathbf{X}^{(k)}\mathbf{w}_k$ (component construction)
  $\mathbf{X}^{(k+1)}$=$\mathbf{X}^{(k)}$ - $\mathbf{P}_{[\mathbf{t}_k]}(\mathbf{X}^{(k)})$ (deflation step)
**end for**

Les composantes PLS $(t_1, \ldots, t_K)$ construites par l'algorithme sont rassemblées dans la matrice $T \in \mathbb{R}^{n\times K}$, où chaque colonne $k$ de $T$ correspond à $t_k$. En particulier on peut noter que $[T] = [XW]$ où $W = (w_1, \ldots, w_K) \in \mathbb{R}^{p\times K}$ est la matrice contenant les poids $w_k$. La prédiction PLS associée $\hat{Y}_{PLS} := X\hat{\beta}_{PLS}$ est donnée par

$$\hat{Y}_{PLS} = T(T^TT)^{-1}T^TY = P_{[XW]}(Y),$$

où l'exposant $T$ désigne la transoposition. L'estimateur de $\beta$ se calcule ensuite de la manière suivante :

$$\hat{\beta}_{PLS} = \hat{\beta}_W = W(W^T\Sigma W)^{-1}W^T\hat{\sigma}, \tag{3}$$

où $\hat{\sigma} = X^TY/n$ correspond à la covariance empirique entre $X$ et $Y$.

Bien que le principe des Moindres Carrés Partiels (PLS) ait attiré beaucoup d'attention au fil des ans, peu de résultats théoriques ont été obtenus. Entre autres, on peut citer Helland (1990) qui a caractérisé l'espace [W] résultant de l'approche PLS à l'aide de $\Sigma$ et $\sigma$. Cook, Li et Chiaromonte (2010) et Cook, Helland et Su (2013) ont établi un lien entre la régression PLS et les enveloppes. Alors que Chun et Keleş (2010) ont prouvé l'inconsistance de l'estimateur PLS lorsque le nombre de covariables est trop important, Cook et Forzani (2017) ainsi que Cook et Forzani (2019) ont établi - sous des contraintes fortes sur $\beta$ et la matrice de design X - le comportement asymptotique de l'erreur quadratique moyenne de prédiction et ont démontré qu'elle peut tendre vers 0 lorsque le nombre d'observations tend vers l'infini. Par ailleurs, des travaux ont également proposé de prendre en compte des contraintes de parcimonie dans l'algorithme PLS. Nous renvoyons par exemple à Durif et al. (2017) et Alsouki et al. (2023).

# 3 Contribution

Nous résumons ici les résultats obtenus dans Castelli, Gannaz et Marteau (2023). Nous nous concentrons sur le cas de la régression PLS sur une composante, c'est à dire $K = 1$. Nous avons dans ce cadre une formule explicite de l'estimateur PLS :

$$\hat{\beta}_{PLS} = \frac{\hat{\sigma}^T\hat{\sigma}}{\hat{\sigma}^T\Sigma\hat{\sigma}}\hat{\sigma}. \tag{4}$$

Nous établissons une borne non asymptotique sur la perte de prédiction. En désignant par $\hat{\beta}_{PLS}$ l'estimateur PLS avec $K = 1$, nous obtenons le résultat suivant :

**Théorème.** *Soit $\delta \in (0,1)$. Supposons que*

$$\hat{\sigma}^T \Sigma \hat{\sigma} > \mathrm{t}_\delta \, p_n \text{ avec } p_n = \frac{\tau^2}{n} \rho(\Sigma) \, Tr(\Sigma) \tag{5}$$

*où $\mathrm{t}_\delta$ est une constante dépendant uniquement de $\delta$. Alors, avec une probabilité supérieure à $1-\delta$, il existe une constante $C_\delta > 0$, dépendant uniquement de $\delta$, telle que*

$$\frac{1}{n} \|X\hat{\beta}_{PLS} - X\beta\|^2 \leq B(\beta) + C_\delta \frac{\tau^2}{n} \max\left( \frac{\mathrm{Tr}(\Sigma)}{\lambda}, \frac{\rho(\Sigma)\mathrm{Tr}(\Sigma)}{\lambda^2} \right), \tag{6}$$

*avec $B(\beta) = \frac{2}{n} \inf_{v \in [\sigma]} \|X(\beta - v)\|^2$ et*

$$\lambda = \frac{\sigma^T \Sigma \sigma}{\sigma^T \sigma}. \tag{7}$$

$\mathrm{Tr}(.)$ *est l'opérateur de trace sur $\mathbb{R}^{p \times p}$.*

Nous renvoyons à Castelli, Gannaz et Marteau (2023) pour la preuve de ce résultat. Les quantités $t_\delta$ et $C_\delta$ ne sont pas données de façon explicite ici, mais elles peuvent être trouvées dans la référence sus-citée.

Dans Castelli, Gannaz et Marteau (2023), nous étendons ce résultat au cas parcimonieux en utilisant une version sparse $\hat{\beta}_{sPLS}$ de l'algorithme, comprenant une contrainte $\ell_1$ dans le processus d'optimisation des poids $w_k$. En supposant que la matrice de Gram $\Sigma$ satisfait une condition de valeurs propres restreintes, similaire à Bickel, Ritov et Tsybakov (2009), nous établissons que, avec grande probabilité,

$$\frac{1}{n} \|X\tilde{\beta}_{sPLS} - X\beta\|^2 \leq B(\beta) + C \frac{\tau^2 s}{n} \ln(p),$$

où $s$ représente le nombre de coefficients non nuls du premier axe PLS et $C$ est une constante. En particulier, nous retrouvons, à l'exception du terme de biais, le même type de borne que celles obtenues pour la procédure Lasso (nous renvoyons à Tibshirani (1996) et Bickel, Ritov et Tsybakov (2009)). Ce résultat n'est pas détaillé ici.

La condition (5) peut être interprétée comme une condition de ratio signal sur bruit qui doit être suffisamment élevé pour que l'estimateur $\hat{\beta}_{PLS}$ contienne de l'information. En effet, on peut montrer que cette condition revient à supposer que la norme de la composante $t_1$ obtenue par l'algorithme PLS est suffisamment grande. Si cette norme est proche de 0, la regression PLS n'est pas pertinente. Si le signal contenu dans la première composante est plus grand qu'un seuil donné, cela assure la qualité de l'estimation.

Ensuite, la quantité $B(\beta)$ est une mesure du biais induit par l'algorithme. Le deuxième terme de la borne (6) peut être considéré comme un terme de variance. Il mesure essentiellement l'impact du bruit $\varepsilon$ sur l'algorithme PLS. Il s'agit d'un rapport entre la trace de $\Sigma$ et le terme $\lambda$ introduit dans (7). La quantité $\lambda$ correspond à la norme théorique de la première

composante PLS $t_1$. En d'autres termes, le terme $\frac{\mathrm{Tr}(\Sigma)}{\lambda}$ peut être considéré comme l'inverse de l'inertie relative. Il fournit une sorte de rapport signal/bruit inverse qui contrôle la précision de la composante PLS unique. Si ce rapport est proche de 1, la première composante PLS capture la majeure partie de l'inertie des données et nous obtenons un terme de variance avec un taux paramétrique $\tau^2/n$. En revanche, si ce rapport est élevé, on ne peut pas s'attendre à obtenir des résultats précis pour la PLS à une seule composante : la quantité de signal capturée dans la première composante n'est pas assez importante.

Cook et Forzani (2016) ont établi un résultat similaire. Leur résultat s'exprime en fonction de $\frac{1}{\sigma^T\sigma}$ qui, sous leurs hypothèses, est équivalent à $1/\lambda$ (voir leur section 3.2 page 12). Notre résultat apporte deux principales contributions par rapport à leur travail :

- Nous ne supposons pas, contrairement à Cook et Forzani (2016), que $\beta$ est colinéaire à la première composante $t_1$ avec notamment une hypothèse d'inversibilité sur la matrice $\Sigma$. Ceci induit notamment un terme de biais dans l'étude de la qualité de prédiction. Nous nous plaçons donc dans un cadre plus général.
- Notre étude est non asymptotique. En effet Cook et Forzani (2016) ont établi que la prédiction de l'estimateur $\hat{\beta}_{PLS}$ est asymptotiquement de l'ordre de $1/n$, mais nous montrons que ce résultat est valable en grande probabilité.

Remarquons enfin que nous considérons des covariables $X$ déterministes, tandis que Cook et Forzani (2016) considèrent des covariables $X$ gaussiennes. Ceci implique notamment que notre condition (5) s'exprime en fonction de $X$ (via $\Sigma$ et $\hat{\sigma}$) et que notre borne fasse intervenir le ratio $\frac{\mathrm{Tr}(\Sigma)}{\lambda}$. Nous pensons que le fait de conserver $X$ dans l'étude permet de mieux comprendre les mécanismes de la régression PLS, notamment par la condition de borne inférieure de la norme de la composante et la borne pouvant s'interpréter comme l'inverse de l'inertie relative de l'axe.

# Conclusion

Notre travail concerne l'estimateur PLS sur une composante. Nous fournissons des bornes non-asymptotiques avec covariables fixes pour la qualité de prédiction de cet estimateur. Ces bornes sont obtenues sous la condition que la quantité de signal de cette composante est suffisamment importante.

Notre résultat peut être interprété comme une décomposition biais-variance reflétant l'importance de la prise en compte du biais dans nos hypothèses. Nous explicitons la borne de la prédiction, avec une interprétation en inertie relative inverse. Ceci permet de mettre en lumière le rôle des composantes dans la régression PLS : plus la composante PLS explique la variance des covariables $X$, meilleure sera la prédiction.

La principale perspective de ce travail est d'étendre les résultats obtenus au cadre multidimensionnel, c'est-à-dire au cas général où $1 \leq K \leq p$. Des résultats asymptotiques quant à la qualité de prédiction dans le cas multidimensionnel ont été obtenus par Cook

et Forzani (2019). Notre objectif est de fournir des résultats non asymptotiques, avec des bornes explicites, et des hypothèses peu restrictives.

# Bibliographie


Alsouki L., Duval L., Marteau C., El Haddad R. and Wahl F. (2023) Dual-sPLS: a family of Dual Sparse Partial Least Squares regressions for feature selection and prediction with tunable sparsity; evaluation on simulated and near-infrared (NIR) spectra, *Chemometrics and Intelligent Laboratory system*, 237, 104813.

Bickel, Peter J. and Ritov, Ya'acov and Tsybakov, Alexandre B. (2009) Simultaneous analysis of lasso and Dantzig selector *The Annals of Statistics*, 37, 1705–1732.

Castelli L., Gannaz I. and Marteau M. (2023) A non asymptotic analysis of the single component PLS regression, arXiv preprint arXiv:2310.10115.

Chun H. and Keles S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 3-25.

Cook R.D., Li B., and Chiaromonte F. (2010) Envelope models for parsimonious and efficient multivariate linear regression, *Statistica Sinica*, 20(3), 927-960.

Cook, R. D., Helland I. and Su Z. (2013) Envelopes and partial least squares regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 75(5), 851-877.

Cook, R. D. and Forzani L. (2017) Big data and partial least-squares prediction, *Canadian Journal of Statistics*, 46(1), 62-78.

Cook, R.D. and Forzani L. (2019) Partial least squares prediction in high-dimensional regression, *The Annals of Statistics*, 47(2), 884-908..

Delaigle A. and Hall P. (2012) Methodology and theory for partial least squares applied to functional data, *The Annals of Statistics*, 40(1), 322-352.

Durif G., Modolo L., Michaelsson J., Mold J.E., Lambert-Lacroix S. and Picard F. (2017) High dimensional classification with combined adaptive sparse PLS and logistic regression, *Bioinformatics*, 34(3), 485-493.

Garthwaite P.H. (1994) An interpretation of partial least squares, *Journal of the American Statistical Association*, 89(425), 122-127.

Helland I. (1990) Partial least squares regression and statistical models, *Scandinavian journal of statistics*, 17(2), 97-114.

Martens H. and Naes T. (1992) Multivariate calibration, *John Wiley & Sons*.

Mateos-Aparicio G. (2011) Partial least squares (PLS) methods: origins, evolution, and application to social sciences, *Communications in Statistics - Theory and Methods*, 40(13), 2305-2317.

Naik P. and Tsai C. (2000) Partial least squares estimator for single-index models, *Journal*

*of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4), 763-771..

Palermo G., Piraino P. and Zucht H.D. (2009) Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data, *Advances and Applications in Bioinformatics and Chemistry*, 57-70.

Sawatsky M., Clyde M. and Meek F. (2015) Partial least squares regression in the social sciences, *The Quantitative Methods for Psychology*, 11(2), 52-62.

Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Wold S. (1995) Chemometrics; what do we mean with it, and what do we want from it?, *Chemometrics and Intelligent Laboratory Systems*, 30(1), 109-115.

Wold S., Sjöström M. and Eriksson L. (2001), PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130.

Yang T.C., Aucott L.S., Duthie G.G. and Macdonald H.M. (2017) An application of partial least squares for identifying dietary patterns in bone health, *Archives of osteoporosis*, 12, 1-8.

# ICS et sous-espace de Fisher : généralisation à plus de deux groupes

Colombe Becquart[1] & Aurore Archimbaud[2] & Klaus Nordhausen[3] & Anne Ruiz-Gazen[1]

[1] *Toulouse School of Economics, Université de Toulouse Capitole, France, colombe.becquart@tse-fr.eu, anne.ruiz-gazen@tse-fr.eu*
[2] *TBS Business School, France, a.archimbaud@tbs-education.fr*
[3] *Department of Mathematics and Statistics, University of Jyväskylä, Finland, klaus.k.nordhausen@jyu.fi*

**Résumé.** L'analyse de la structure multidimensionnelle des données telle que l'identification de clusters est souvent rendue complexe par le nombre de dimensions à analyser. Lorsque cette structure est contenue dans un sous-espace, il est ainsi bénéfique de réduire la dimension pour se placer dans ce sous-espace d'intérêt. L'analyse en composantes principales (ACP) est la méthode de référence pour la réduction de dimension. Dans cet article, nous étudions une méthode alternative à l'ACP, appelée "Invariant Coordinate Selection" (ICS). Contrairement à l'ACP, ICS ne repose pas sur la maximisation de la variance mais sur la maximisation/minimisation d'un kurtosis généralisé, et n'est pas uniquement invariante par transformation orthogonale des données mais par toute transformation affine. Plus précisément, ICS consiste à comparer deux matrices de dispersion au travers de leur diagonalisation jointe. La réduction de dimension est obtenue en projetant les données sur les vecteurs propres associés aux plus grandes et plus petites valeurs propres d'ICS. Des travaux empiriques ont montré l'efficacité de la méthode dans le cadre du clustering et de la détection d'anomalies. Certaines propriétés théoriques d'ICS sont aussi connues. En particulier, pour un mélange de distributions elliptiques et sous certaines conditions, une sélection des composantes d'ICS permet de retrouver le sous-espace discriminant de Fisher, quel que soit le choix des matrices de dispersion. Toutefois, les conditions sous lesquelles ce résultat général s'applique ne sont explicites que pour des cas particuliers, tels que des mélanges de deux groupes de même matrice de covariance. L'objectif de cet article est d'explorer plus avant le comportement théorique d'ICS dans le cadre d'un mélange de lois gaussiennes de même matrice de covariance pour un nombre de groupes quelconque $k$. Les matrices de dispersion considérées sont la matrice de covariance et la matrice basée sur les moments d'ordre 4. Dans ce contexte de $k \geq 2$ groupes, nous étudions les conditions sous lesquelles ICS "fonctionne", c-à.-d. sous lesquelles les composantes associées aux $k-1$ plus grandes et plus petites valeurs propres engendrent le sous-espace de Fisher qui contient les moyennes des groupes. A partir de calculs théoriques et numériques, nous montrons que pour des groupes suffisamment séparés, ces conditions s'expriment essentiellement en fonction des proportions de chaque groupe et que les valeurs des moyennes des groupes ont peu d'influence.

**Mots-clés.** Clustering, diagonalisation jointe, modèles de mélange, réduction de dimension.

**Abstract.** Analyzing multidimensional data structures, such as identifying clusters, is frequently complexified by the number of dimensions to be analyzed. When the data structure is contained within a subspace, it is therefore beneficial to reduce the dimension to the subspace of interest. Principal component analysis (PCA) is the reference method. In this article, an alternative method to PCA is studied, it is called "Invariant Coordinate Selection" (ICS). Unlike PCA, ICS is not based on variance maximization but on the maximization/minimization of a generalized kurtosis, and it is invariant not only to orthogonal data transformations but to any affine transformation. Specifically, ICS compares two scatter matrices through their joint diagonalization. Dimension reduction is achieved by projecting the data onto the eigenvectors associated with the largest and smallest eigenvalues obtained with ICS. Empirical work has shown the relevance of applying ICS to clustering and anomaly detection. Some theoretical properties of ICS are also known. In particular, for a mixture of elliptical distributions and under certain conditions, a selection of invariant components recovers the Fisher's linear discriminant subspace for any scatter matrices. However, the conditions under which this general result holds are not always explicit. They are derived only for some special cases, such as mixtures of two groups with the same covariance matrix. The purpose of this article is to further explore the theoretical behavior of ICS in the context of a mixture of Gaussian distributions with the same covariance matrix for any number of groups $k$. The scatter matrices under consideration are the covariance matrix and the scatter matrix of fourth moments. In this context of $k \geq 2$ groups, we study the conditions under which ICS "works", i.e. under which the components associated with the $k-1$ largest and smallest eigenvalues span the Fisher's linear discriminant subspace containing the group means. Using theoretical and numerical calculations, we show that for sufficiently separate groups, these conditions are mainly related to the group proportions and that the group means has only a minor influence.

**Keywords.** Clustering, dimension reduction, joint diagonalization, mixture models.

# 1 Introduction

L'analyse multidimensionnelle vise à explorer les relations complexes entre plusieurs variables d'un ensemble de données. Dans ce contexte, des méthodes statistiques telles que la classification non supervisée et la détection d'anomalies sont utiles dans une multitude de domaines tels que l'industrie, le marketing ou l'économie. Cependant, lorsque la dimension des données est grande (en terme de nombre de variables) comparée à la dimension de la structure d'intérêt, l'identification de cette structure est rendue plus difficile. C'est le cas par exemple si le nombre de groupes en classification non supervisée est petit comparé au nombre de variables. L'espace contenant les centres de groupes est alors de dimension inférieure à celle de l'espace où se trouvent les données. Dans de tels cas, la réduction de la dimension peut s'avérer bénéfique pour se concentrer sur le sous-espace contenant la structure des données.

Invariant Coordinate Selection (ICS) est une méthode de réduction de dimension intro-

duite dans l'article de Tyler et al. (2009). L'article de Caussinus et Ruiz (1990) sur l'Analyse en Composantes Principales Généralisée peut être considéré comme un travail précurseur d'ICS. ICS se rapproche de l'analyse en composantes principales (ACP) dans la mesure où il s'agit d'une transformation ayant pour objectif de réduire la dimension tout en préservant au maximum la structure des données. A la différence de l'ACP, ICS ne se base pas sur la maximisation de la variance, mais sur l'optimisation d'un kurtosis généralisé à travers la diagonalisation jointe de deux matrices de dispersion. ICS se distingue également de l'ACP par les deux propriétés suivantes. Premièrement, les composantes d'ICS sont invariantes (au signe et permutation près) pour toute transformation affine, contrairement aux composantes de l'ACP qui ne sont invariantes que pour des transformations orthogonales. Deuxièmement, cette méthode permet de retrouver le sous-espace discriminant de Fisher lorsque les données suivent un mélange de distributions elliptiques et sous certaines conditions qui ne sont pas faciles à expliciter en dehors de cas particuliers tels qu'un mélange de deux lois gaussiennes (Tyler et al., 2009).

Dans le cadre de la détection d'anomalies, ICS a été étudié sur un plan théorique et empirique par Archimbaud, Nordhausen, et Ruiz-Gazen (2018). Dans le cadre de la classification non-supervisée, les propriétés théoriques d'ICS sont connues dans le cas de deux groupes (Tyler et al., 2009; voir aussi Peña et Prieto, 2001). Pour plus de deux groupes, Alfons et al. (2024) proposent une étude empirique.

Après avoir rappelé le principe d'ICS dans la section 2.1, nous décrivons dans la section 2.2 le comportement bien connu de la méthode dans le cas d'un mélange de deux gaussiennes. L'objectif de notre présentation est d'étudier théoriquement et numériquement le comportement d'ICS lorsqu'il y a plus que deux groupes. L'étude se concentre sur un mélange de lois gaussiennes de même matrice de covariance et sur ICS avec la matrice de covariance et la matrice basée sur les moments d'ordre 4. Cette étude est présentée de manière théorique dans la section 3 et de manière empirique dans la section 4.

# 2 Invariant Coordinate Selection

Soit $Y$ un vecteur aléatoire de dimension $p$, et de fonction de répartition $F_Y$. Une matrice de dispersion de $Y$, notée $V(F_Y)$, est une matrice symétrique définie positive de dimension $p \times p$ et affine équivariante, c-à.-d $V(F_{AY+b}) = AV(F_Y)A'$ pour toute matrice $A$ non singulière de dimension $p \times p$ et pour tout $b \in \mathbb{R}^p$. On note $\mathcal{P}_p$ l'ensemble de toutes les matrices symétriques définies positives d'ordre $p$ et on simplifie l'écriture de $V(F_Y)$ en $V$ lorsque le contexte est clair.

## 2.1 Principe général d'ICS

Le principe d'ICS est de comparer deux matrices de dispersion affines équivariantes, notées $V_1$ et $V_2$. Si la distribution des données est elliptique, alors toutes les matrices de dispersion affines équivariantes sont proportionnelles. Dans le cas contraire, $V_1$ et $V_2$ n'ont pas la même forme et sont donc différentes sur certaines dimensions. Identifier les directions dans lesquelles

elles diffèrent permet de retrouver les déviations par rapport à une distribution elliptique, et donc de révéler la structure des données. Pour effectuer cette comparaison, ICS repose sur la diagonalisation jointe de $V_1$ et $V_2$ :

$$H'V_1H = D_1$$
$$H'V_2H = D_2$$

où $V_1, V_2 \in \mathcal{P}_p$, $D_1$ et $D_2$ sont des matrices diagonales telles que $D_1^{-1}D_2 = diag(\rho_1, \ldots, \rho_p)$, $\rho_1, \ldots, \rho_p$ étant les valeurs propres de $V_1^{-1}V_2$ triées dans un ordre décroissant, et $H = (h_1, \ldots, h_p)$ est une matrice contenant les vecteurs propres correspondant.

La transformation $Z = H'Y$ présente deux propriétés intéressantes. Premièrement, les variables obtenues sont invariantes sous transformation affine (voir théorèmes 1 et 2 de Tyler et al., 2009). Deuxièmement, elle permet de retrouver le sous-espace discriminant de Fisher lorsque les données suivent un mélange de distributions elliptiques et sous certaines conditions (voir théorèmes 3 et 4 de Tyler et al., 2009). En particulier, Tyler et al. (2009) s'intéressent au cas où le mélange est formé de distributions de même matrice de dispersion, mais dont les centres et/ou fonctions de densité peuvent différer. Les centres sont de dimension $p$ et on suppose qu'ils génèrent un sous-espace de dimension $q$ telle que $0 < q < p$. Le théorème 4 de Tyler et al. (2009) nous apprend que dans ce cas il y a au moins une valeur propre issue de la diagonalisation jointe des deux matrices de dispersion qui est de multiplicité supérieure ou égale à $p - q$. Lorsqu'une valeur propre a une multiplicité égale à $p - q$, et non supérieure, le sous-espace engendré par les vecteurs propres associés aux autres valeurs propres qui sont les plus grandes ou les plus petites, est le sous-espace discriminant de Fisher. Pour réduire la dimension, il suffit alors de projeter les données sur les vecteurs propres associés aux plus grandes et plus petites valeurs propres d'ICS.

## 2.2 Cas d'un mélange de 2 groupes

Lorsque la distribution est un mélange de deux distributions gaussiennes, le cas où la multiplicité est supérieure à $p - q$ est celui où ICS ne fonctionne pas : toutes les valeurs propres sont égales. En prenant comme matrice de dispersion la matrice de covariance et la matrice de dispersion basée sur les quatrièmes moments, il est montré que le cas décrit précédemment se produit lorsque la proportion d'un groupe est de $(3-\sqrt{3})/6$, soit environ 21% (Tyler et al., 2009). Si la proportion d'un groupe est inférieure à ce seuil, il y a une valeur propre strictement supérieure aux autres qui sont égales entre elles. Inversement, si les deux groupes ont une proportion supérieure à ce seuil, il y aura une valeur propre strictement inférieure aux autres qui seront égales entre elles. Le cas d'un mélange de deux distributions gaussiennes a été approfondi par Archimbaud, Nordhausen, et Ruiz-Gazen (2018) qui recommandent dans un contexte de détection d'anomalies d'utiliser la matrice de covariance, notée $COV$, pour $V_1$ et la matrice basée sur les moments d'ordre 4, notée $COV_4$, pour $V_2$ :

$$COV(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))']$$

$$\mathrm{COV}_4(Y) = \frac{1}{p+2}\mathbb{E}[d^2(Y-\mathbb{E}(Y))(Y-\mathbb{E}(Y))']$$

où $d^2$ est la distance de Mahalanobis au carré :

$$d^2 = (Y-\mathbb{E}(Y))'COV^{-1}(Y)(Y-\mathbb{E}(Y))$$

Avec cette paire, les auteurs démontrent les bonnes performances d'ICS dans un contexte de détection d'anomalies où l'on considère qu'une large proportion des données suit une loi gaussienne tandis qu'une plus faible proportion (les données atypiques) suit une loi gaussienne de moyenne différente. Grâce aux propriétés d'ICS, il suffit dans ce cas de sélectionner la première composante. De plus, les cas où ICS ne fonctionne pas sont assez rares puisqu'ils correspondent uniquement au cas où le groupe des atypiques est en proportion d'environ 21%. Cependant, limiter l'étude théorique à seulement deux groupes est restrictif et nous cherchons à savoir si les résultats obtenus sont généralisables à $k$ groupes. La suite de l'article examine le comportement d'ICS dans un contexte de mélange de $k \geq 2$ distributions gaussiennes, en utilisant $COV$ et $COV_4$.

# 3 Généralisation à $k$ groupes

Nous cherchons à obtenir des résultats théoriques lorsqu'il y a $k \geq 2$ groupes, en se concentrant sur un mélange de distributions gaussiennes de même matrice de covariance. Les matrices de dispersion utilisées sont $COV$ et $COV_4$. Comme dans la section précédente, l'objectif est de trouver les conditions pour lesquelles ICS "fonctionne", c'est-à-dire les cas où toutes les valeurs propres ne sont pas égales. Pour cela, on réécrit le modèle de manière à expliciter la forme de $COV^{-1}\mathrm{COV}_4$ et les valeurs propres associées. Devant les défis posés par les calculs théoriques, nous avons choisi d'utiliser des approximations numériques pour certaines étapes et de visualiser notamment les approximations numériques des valeurs propres.

## 3.1 Simplification du modèle

Cette sous-section reprend la transformation introduite par Tyler et al. (2009) dans la preuve de son théorème 4, afin de simplifier le modèle sans perte de généralité. $Y$ est un vecteur aléatoire suivant un mélange de $k$ distributions gaussiennes de même matrice de covariance :

$$Y \sim \sum_{j=1}^{k} \alpha_j N_p(\mu_j, \Gamma)$$

où $\alpha_j > 0$ pour $j = 1, \ldots, k$ et $\sum_{j=1}^{k} \alpha_j = 1$, $\mu_j \in \mathbb{R}_p$ pour $j = 1, \ldots, k$, et $\Gamma \in \mathcal{P}_p$.

Soit $M = (\mu_1, ..., \mu_k)$ la matrice contenant les centres de $Y$. On pose $M_0 = \Gamma^{-\frac{1}{2}}(M - \mu_k \mathbf{1}'_k)$ avec $\mathbf{1}_k$ un vecteur de 1 de dimension $k$. On note $q$ le rang de la matrice $M_0$. La décomposition QR de $M_0$ est :

$$M_0 = PT = P \begin{pmatrix} T_u & 0 \\ 0 & 0 \end{pmatrix}$$

où P est une matrice orthogonale, $T = [t_1, \ldots, t_k]$ et $T_u$ est une matrice triangulaire supérieure de dimension $k - 1 \geq 1$ telle que les $k - 1 - q \geq 0$ dernières lignes sont nulles.
Soit $X = P'\Gamma^{-\frac{1}{2}}(Y - \mu_k \mathbf{1}'_k)$. X est un mélange de $k$ distributions normales de centres $t_1, \ldots, t_k$ et de matrice de covariance l'identité, avec des poids $\alpha_1, \ldots, \alpha_k$. Grâce à la propriété d'invariance affine d'ICS, on considère sans perte de généralité le modèle suivant :

$$X \sim \sum_{j=1}^{k} \alpha_j N_p(t_j, I_p)$$

## 3.2 Calcul des valeurs propres

A partir du modèle précédent, il est possible d'écrire les matrices $COV^{-1}$ et $\mathrm{COV}_4$ sous les formes suivantes :

$$COV(X)^{-1} = \begin{bmatrix} B & 0 \\ 0 & I_{p-k+1} \end{bmatrix}, \qquad COV_4(X) = \begin{bmatrix} A & 0 \\ 0 & I_{p-k+1} \end{bmatrix},$$

où $A = [a_{ij}]$ et $B = [b_{ij}]$ sont deux matrices $(k-1) \times (k-1)$.

Nous n'explicitons pas les termes de la matrices $B$ qui dans la suite seront calculés numériquement. Pour la matrice $A$, nous avons:

$$\begin{aligned} a_{mn} &= \frac{1}{p+2}\mathbb{E}\left[\sum_{i=1}^{k-1}\sum_{j=1}^{k-1} x_m^{\mathrm{c}} x_n^{\mathrm{c}} x_i^{\mathrm{c}} x_j^{\mathrm{c}} b_{ij} + x_m^{\mathrm{c}} x_n^{\mathrm{c}} \sum_{i=k}^{p} (x_i^{\mathrm{c}})^2\right] \\ &= \frac{1}{p+2}\left[\sum_{i=1}^{k-1}\sum_{j=1}^{k-1} b_{ij}\mathbb{E}[x_m^{\mathrm{c}} x_n^{\mathrm{c}} x_i^{\mathrm{c}} x_j^{\mathrm{c}}] + (p-k+1)\mathbb{E}[x_m^{\mathrm{c}} x_n^{\mathrm{c}}]\right] \end{aligned}$$

pour $m, n = 1, \ldots, k-1$ et avec $(x_1^{\mathrm{c}}, \ldots, x_p^{\mathrm{c}})^T = X - \mathbb{E}(X)$. Les moments dans l'expression des termes $a_{mn}$ sont calculables explicitement mais pas détaillés ici.

Le produit des deux matrices de dispersion donne :

$$COV^{-1}\mathrm{COV}_4 = \begin{bmatrix} BA & 0 \\ 0 & I_{p-k+1} \end{bmatrix}$$

et nous cherchons les valeurs propres de ce produit. Du fait de la structure du produit, il y a au moins $p - k + 1$ valeurs propres égales à 1. Le calcul des autres valeurs propres conduit à des expressions compliquées, fonctions des termes $b_{ij}$ de la matrice $B$. Pour les calculer, nous avons recours à des calculs numériques notamment pour l'inverse de la matrice de covariance ainsi que pour les valeurs propres et vecteurs propres de $COV^{-1}\mathrm{COV}_4$.

# 4 Résultats empiriques et conclusion

L'objectif des calculs numériques est de comprendre le comportement des valeurs propres de $COV^{-1}\text{COV}_4$ pour des mélanges de lois gaussiennes de plus de deux groupes en fonction de la configuration des groupes. Avec la simplification du modèle présentée dans la section 3.1, les matrices de covariances sont égales à l'identité et la configuration des groupes dans le mélange dépend des proportions des différents groupes et de leurs moyennes. L'un des objectifs est de détecter les cas où toutes les valeurs propres sont égales, c'est-à-dire les cas pour lesquels ICS ne fonctionne pas. Dans notre cas, avec les matrices de dispersion choisies, les valeurs propres égales valent 1. On rappelle que dans le cas de deux groupes, le fait qu'ICS ne fonctionne pas ne dépend que de la proportion des groupes et pas des moyennes des groupes.

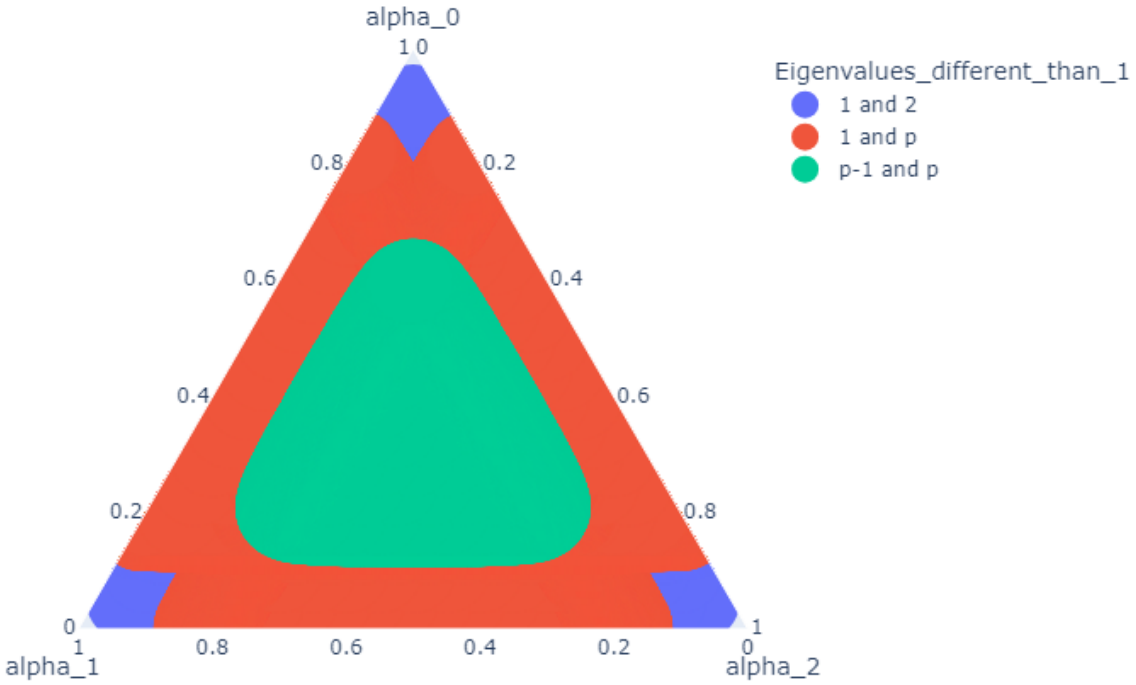



Figure 1: Diagramme ternaire représentant les valeurs propres de $COV^{-1}\text{COV}_4$ selon les proportions des 3 groupes notées alpha_0, alpha_1 et alpha_2. Les moyennes de groupes utilisées pour la génération de ce graphique sont $t_1 = (200, 0, 0, \ldots, 0)$, $t_2 = (400, 100, 0, \ldots, 0)$, $t_3 = (0, 0, \ldots, 0)$, et $p = 13$.

Dans un premier temps, nous étudions des scénarios où les groupes sont bien séparés. Une fois les moyennes de groupes choisies, une grille de proportions de groupes est générée et, pour chaque combinaison, les valeurs propres de $COV^{-1}\text{COV}_4$ sont calculées. Nous avons

répété ce protocole en faisant varier le nombre de groupes de 3 à 11. La Figure 1 illustre les résultats que nous avons obtenus avec 3 groupes et $q = 2$. On rappelle que les valeurs propres sont triées dans un ordre décroissant. Lorsqu'il y a deux groupes de faible proportion, il y a deux valeurs propres supérieures à 1 (zone bleue sur la Figure 1). Dans le cas où il n'y a qu'un groupe de faible proportion, il y a une valeur propre supérieure à 1 et une inférieure à 1 (zone rouge sur la Figure 1). Enfin, dans le cas où il n'y a aucun groupe de faible proportion, les deux valeurs propres différentes de 1 sont inférieures à 1 (zone verte sur la Figure 1). Même s'il n'est pas facile de déterminer précisément le seuil à dépasser pour considérer qu'un groupe à une "faible" proportion, on peut dire que ce seuil est strictement inférieur à 21% et de l'ordre de 18% pour 3 groupes. De plus, même si le graphique ne montre que des cas où il y a 2 valeurs propres différentes de 1, il est possible que les frontières entre les zones bleue, rouge et verte correspondent au cas où ICS ne fonctionne pas. De manière intéressante, nous avons obtenu des résultats très similaires en faisant varier les moyennes de groupes mais aussi des résultats comparables quand on augmente le nombre de groupes. Ainsi, nous faisons la conjecture que, pour des groupes bien séparés, les conditions de bon fonctionnement d'ICS s'expriment d'avantage en fonction des proportions de chaque groupe qu'en fonction des valeurs des moyennes des groupes.

Pour approfondir l'analyse, nous souhaitons utiliser ces calculs numériques pour étudier ICS dans d'autres situations. D'abord, nous examinerons plus en détail les cas impliquant plus de trois groupes dans le mélange de distributions gaussiennes. Ces cas sont néanmoins plus difficile à représenter graphiquement. De plus, nous étudierons les cas où on observe de la colinéarité entre les moyennes des groupes, c-à.-d. les cas où $0 < q < k - 1$.

# Bibliographie


Alfons, A., Archimbaud, A., Nordhausen, K. and Ruiz-Gazen, A. *Tandem clustering with invariant coordinate selection.* arXiv:2212.06108 [stat].

Archimbaud, A., Nordhausen, K. and Ruiz-Gazen, A. (2018). ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, 128, pp. 184-199.

Caussinus, H. and Ruiz, A. (1990), Interesting Projections of Multidimensional Data by Means of Generalized Principal Component Analyses. In: Momirović, K., Mildner, V. (Eds) *Compstat* (pp. 121–126). Physica-Verlag HD.

Peña, D. and Prieto, F. J. (2001). Cluster identification using projections. *Journal of the American Statistical Association*, 96(456), 1433-1445.

Tyler, D. E., Critchley, F., Dümbgen, L. and Oja, H. (2009). Invariant co-ordinate selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(3), pp. 549-592.

# Importance sampling

# Importance sampling for online variational learning in state-space models

Mathis Chagneux [1], Pierre Gloaguen [2], Sylvain Le Corff [3], Jimmy Olsson [4]

[1] *Institut Polytechnique de Paris, France, mathis.chagneux@telecom-paris.fr*
[2] *Université Bretagne Sud, France, pierre.gloaguen@univ-ubs.fr*
[3] *Sorbonne Université, France, sylvain.le_ corff@sorbonne-universite.fr*
[4] *KTH Royal Institute of Technology, Sweden, jimmyol@kth.se*

**Résumé.** Dans ce travail, nous considérons le problème de l'apprentissage en ligne dans les modèles espace d'états, c'est à dire un cadre où les observations dépendent d'états cachés, eux mêmes supposés issus un processus de Markov. Notre objectif est d'apprendre la distribution de lissage, *i.e.* la distribution a posteriori de états cachés conditionnellement aux observations. Nous nous intéressons au cadre de l'apprentissage en ligne, c'est à dire où l'actualisation de la loi se fait à l'arrivée de chaque nouvelle observation, et chaque observation n'est vue qu'une seule fois. Nous proposons un nouvel algorithme pour estimer en ligne la distribution de lissage dans un cadre variationel. Cet algorithme repose sur une estimation en ligne efficace de la fonction de coût classique en inférence variationelle, l'*evidence lower bound* (ELBO), ainsi que de son gradient. Nous mimons ensuite les idées du maximum de vraisemblance récursif pour l'apprentissage en ligne. Nous montrons comment on peut exploiter i) la structure de la vraie loi a posteriori ciblée ii) les idées des approches de Monte Carlo séquentiel et iii) les astuces de paramétrisation des approches variationnelles récentes pour apprendre efficacement une loi a posteriori en grande dimension.

**Mots-clés.** Apprentissage en ligne ; Inférence variationnelle séquentielle ; Modèles à espace d'état ; Apprentissage pour les séries temporelles

**Abstract.** In this work, we consider the problem of online learning in state-space models, *i.e.* when the observations are time-series depending on hidden states, themselves assumed to arise from a Markov process. Our objective is to learn the smoothing distribution, *i.e.* the a posteriori distribution of hidden states conditional on the observations. We are interested in the online learning framework, *i.e.* where the updating of the distribution takes place upon the arrival of each new observation, and each observation is seen only once. We propose a new algorithm for online estimation of the smoothing distribution in a variational framework. This algorithm is based on an efficient online estimation of the classical cost function in variational inference, the *evidence lower bound* (ELBO), as well as its gradient. We then mimic the ideas of recursive maximum likelihood for online learning. We show how we can exploit i) the structure of the true targeted a posteriori law ii) the ideas of sequential Monte Carlo approaches and iii) the parameterization tricks of recent variational approaches to efficiently learn a high-dimensional a posteriori law.

**Keywords.** Online learning ; Sequential variational inference ; State-space models ; Time-series learning

# 1 Modèle et objectif d'inférence

Soit un modèle de Markov caché (HMM) où le processus caché, à valeurs dans $\mathbb{R}^{d_x}$, est noté $(X_t)t \geqslant 0$. On suppose que la distribution de $X_0$ admet une densité $\chi$ par rapport à la mesure de Lebesgue $\mu$ et pour tout $t \geqslant 0$, la distribution conditionnelle de $X_{t+1}$ conditionnellement à $X_{0:t}$ admet une densité $m_t(X_t, \cdot)$. Dans un HMM, on suppose que cet état est partiellement observé à travers un processus d'observation $(Y_t)t \geqslant 0$ prenant des valeurs dans $\mathbb{R}^{d_y}$. Les observations $Y_{0:t}$ sont supposées indépendantes conditionnellement à $X_{0:t}$ et, pour tout $t \geqslant 0$, la distribution de $Y_t$ sachant $X_{0:t}$ dépend de $X_t$ uniquement. On suppose que cette distribution admet une densité $g_t(X_t, \cdot)$ par rapport à la mesure de Lebesgue. L'ensemble du modèle est alors défini par la distribution conjointe des états cachés et des observations :

$$p_{0:t}(x_{0:t}, y_{0:t}) = \prod_{s=0}^{t} \ell_s(x_{s-1}, x_s, y_s),$$

où

$$\begin{aligned} \ell_0(x_{-1}, x_0, y_0) &:= \chi(x_0) g_0(x_0, y_0) \\ \ell_s(x_{s-1}, x_s, y_s) &:= m_s(x_{s-1}, x_s) g_s(x_{s-1}, y_s) \qquad \text{for } s \geqslant 1 \ . \end{aligned}$$

## 1.1 Lissage dans les HMMs

Un des objectifs d'inférence classique dans les HMM est d'estimer la valeur prise par les états cachés, conditionnellement aux observations. Formellement, cela consiste à estimer la distribution de lissage, c'est-à-dire la loi conditionnelle de $X_{0:t}$ sachant $Y_{0:t}$. Dans notre cadre (où tout est défini via des densités), cette distribution admet également une densité donnée par

$$\phi_{0:t}(x_{0:t}) \propto p_{0:t}(x_{0:t}, y_{0:t}) \ .$$

La marginale au temps $t$ de cette distribution conjointe est appelée *distribution de filtrage* au temps $t$, et sa densité par rapport à la mesure de Lebesgue s'écrit $\phi_t$. On montre facilement que cette densité de lissage peut se factoriser de la manière suivant :

$$\phi_{0:t}(x_{0:t}) = \phi_t(x_t) \prod_{s=1}^{t} b_{s-1|s}(x_s, x_{s-1}) \ . \tag{1}$$

où, pour $1 \leqslant s \leqslant t$, les noyaux

$$b_{s-1|s}(x_s, x_{s-1}) = \frac{m_s(x_{s-1}, x_s)\phi_{s-1}(x_{s-1})}{\int m_s(x_{s-1}, x_s)\phi_{s-1}(x_{s-1}) \, \mathrm{d}x_{s-1}}, \tag{2}$$

sont appelés noyaux *backward*. À $x_s$ fixé, ce noyau donne la densité conditionnelle de $X_{s-1}$ étant donnés $(X_s = x_s, Y_{0:s-1} = y_{0:s-1})$. Une remarque importante est que la factorisation donnée dans l'équation (2) met en lumière le caractère markovien de la distribution de lissage. La loi de lissage n'a généralement pas d'expression explicite en raison de l'impossibilité de calculer les lois de filtrage. La section suivante décrit comment on estime la loi de lissage dans un cadre variationel.

## 1.2 Inférence variationelle séquentielle *backward*

Dans les approches variationnelles, la distribution de lissage $\phi_{0:t}$ est approximée en choisissant un candidat dans une famille paramétrique $\{q^\lambda_{0:t}\}_{\lambda\in\Lambda}$, appelée famille variationnelle, où $\Lambda$ est un ensemble de paramètres. Un point critique réside donc dans le choix de la forme de la famille variationnelle.

Les approches variationelle classiques [Blei et al., 2017] se basent sur la famille dite de champ moyen, où l'on suppose que :

$$q^\lambda_{0:t}(x_{0:t}) = \prod_{s=0}^{t} q^\lambda_s(x_s, x_{s-1}) \ .$$

Cette famille où tous les états cachés sont donc indépendants a posteriori ne respecte pas la structure markovienne de la distribution de lissage. Par conséquent, la plupart des travaux d'inférence variationelle séquentielle imposent une structure à la famille variationnelle via une décomposition factorisée de $q^\lambda_{0:t}$ sur $x_{0:t}$. Une contrepartie variationnelle de (1), introduite dans les travaux de [Campbell et al., 2021], consiste à définir

$$q^\lambda_{0:t}(x_{0:t}) = q^\lambda_t(x_t) \prod_{s=1}^{t} q^\lambda_{s-1|s}(x_s, x_{s-1}) \ , \tag{3}$$

où les $q^\lambda_t$ (resp. $q^\lambda_{s-1|s}(x_s, \cdot)$) sont des densités de probabilités, que l'on choisit, qui dépendront de $Y_{0:t}$ (resp. $Y_{0:s-1}$). L'un des principaux avantages de cette factorisation est qu'elle respecte les véritables dépendances induites dans (2). En outre, récemment, [Chagneux et al., 2024a] ont établi une borne supérieure sur l'erreur lorsque l'on approche des espérances par rapport à la distribution de lissage, par des espérances par rapport aux distributions variationnelles satisfaisant cette factorisation.

Au sein de cette famille variationnelle, on choisira le meilleur $\lambda$ au sens de l'*evidence lower bound* (ELBO), à savoir le $\lambda$ qui maximise :

$$\mathcal{L}^\lambda_t = \mathbb{E}_{q^\lambda_{0:t}}\left[\log \frac{p_{0:t}(X_{0:t}, Y_{0:t})}{q^\lambda_{0:t}(X_{0:t})}\right] \ . \tag{4}$$

Cette maximisation se fait généralement via des algorithmes de gradient qui nécessitent donc de calculer le gradient de l'ELBO. Les sections suivantes décrivent un nouvel algorithme permettant d'approximer ce gradient de manière récursive.

# 2 Recursions pour le calcul du gradient de l'ELBO

On définit [1]

$$\tilde{f}^\lambda_t(x_{t-1}, x_t) = \begin{cases} \log \ell_0(x_{-1}, x_0, y_0) & \text{if } t = 0, \\ \log \frac{\ell_t(x_{t-1}, x_t, y_t)}{q^\lambda_{t-1|t}(x_t, x_{t-1})} & \text{if } t > 0 \end{cases} \tag{5}$$

et $f^\lambda_{0:t}(x_{0:t}) = \sum_{s=0}^{t} \tilde{f}^\lambda_s(x_{s-1}, x_s)$. On remarque directement qu'avec ces notations

$$\mathcal{L}^\lambda_t = \mathbb{E}_{q^\lambda_{0:t}}\left[f^\lambda_{0:t}(X_{0:t}) - \log q^\lambda_t(X_t)\right].$$

Notre méthode d'estimation de $\lambda$ se base sur les résultats suivant [2] :

1. La dépendance de chaque terme $\tilde{f}^\lambda_t$ en $y_t$ is ommise pour alléger les notations.
2. Dans la suite, tous les gradients sont calculés par rapport à $\lambda$

**Proposition 1.** *Pour l'ELBO et son gradient, on a :*

$$\mathcal{L}_t^\lambda = \mathbb{E}_{q_t^\lambda}\left[H_t^\lambda(X_t)\right] - \mathbb{E}_{q_t^\lambda}\left[\log q_t^\lambda(X_t)\right] \tag{6}$$

$$\nabla\mathcal{L}_t^\lambda = \mathbb{E}_{q_t^\lambda}\left[\{\nabla \log q_t^\lambda \cdot H_t^\lambda\}(X_t) + G_t^\lambda(X_t)\right], \tag{7}$$

*où $H_t^\lambda(x_t)$ est une fonction de $\mathbb{R}^{d_x}$ dans $\mathbb{R}$ satisfaisant la recursion*

$$H_t^\lambda(x_t) = \mathbb{E}_{q_{t-1|t}^\lambda}\left[H_{t-1}^\lambda(X_{t-1}) + \tilde{f}_t^\lambda(X_{t-1}, x_t)\right], \tag{8}$$

*avec $H_0^\lambda(x_0) = \tilde{f}_0^\lambda(x_{-1}, x_0)$, et où $G_t^\lambda(x_t)$ est une fonction de $\mathbb{R}^{d_x}$ dans $\Lambda$ satisfaisant $G_0^\lambda(x_0) = 0$ et*

$$G_t^\lambda(x_t) = \mathbb{E}_{q_{t-1|t}^\lambda}\left[G_{t-1}^\lambda(X_{t-1}) + \nabla \log q_{t-1|t}^\lambda(x_t, X_{t-1})\left(H_{t-1}^\lambda(x_{t-1}) + \tilde{f}_t^\lambda(x_{t-1}, x_t)\right)\right]. \tag{9}$$

*Démonstration.* Voir l'Annexe A. □

**Intérêt de la Proposition 1** Pour une séquence fixe de longueur $t$, le calcul récursif de $\nabla\mathcal{L}_t^\lambda$ consiste donc à (i) calculer de manière récursive $\nabla H_t^\lambda(x_t)$ de 0 à $t$ en utilisant les récursions (8) et (9) et (ii) calculer l'espérance finale grâce à l'équation (7). Cette espérance est prise relativement à la distribution variationnelle choisie $q_t^\lambda$, selon laquelle on sait simuler facilement. Ainsi, on pourra utiliser une approximation Monte Carlo pour l'estimer. Il reste à approximer récursivement les fonctions intermédiaires $H_t^\lambda(x_t)$ et $G_t^\lambda(x_t)$, impliquant des espérances conditionnelles par rapport aux noyaux $q_{t-1|t}^\lambda$. Notez que ces noyaux visent à approximer les vrais noyaux *backward* $b_{t-1|t}$, qui dépendent uniquement des observations $Y_{0:t-1}$, c'est-à-dire seulement du passé. Les récursions proposées sont donc adaptées à l'apprentissage en ligne.

# 3 Apprentissage de $\lambda$ en ligne

Dans l'apprentissage en ligne, nous visons, à chaque paas de temps $t$, à actualiser notre estimation $\hat{\lambda}_t$ de $\lambda$ (obtenue avec les observations $y_{0:t-1}$) à l'aide l'observation $y_t$.

**Initialisation.** À partir d'une estimation initiale $\hat{\lambda}_0$, on simule un $N$-échantillon

$$\{\xi_0^j\}_{1\leq j\leq N} \overset{i.i.d}{\sim} q_0^{\hat{\lambda}_0},$$

et on pose

$$\hat{H}_0^{\hat{\lambda}_0,j} = H_0^{\hat{\lambda}_0}(\xi_0^j),$$
$$\hat{G}_0^{\lambda,j} = G_0^{\hat{\lambda}_0}(\xi_0^j)$$

Il est important de remarquer ici que :

— L'observation $y_0$ intervient dans i) la simulation du $N$-échantillon (la dépendance de $q_0^\lambda$ en $y_0$ étant implicite par souci de notation et ii) dans le calcul de $H_0^{\hat{\lambda}_0}(\xi_0^j)$.

— Les fonctions $H_0^{\hat{\lambda}_0}$ et $G_0^{\hat{\lambda}_0}$ sont simplement évaluées [3] sur un support fini. C'est ce support fini qui permettra la propagation récursive de ces statistiques.

Le premier gradient est donc approché par :

$$\widehat{\nabla}\mathcal{L}_0^{\hat{\lambda}_0} = \frac{1}{N}\sum_{i=1}^{N} \nabla \log q_0^{\hat{\lambda}_0}(\xi_0^i) \cdot \hat{H}_0^{\hat{\lambda}_0,i} + \hat{G}_0^{\hat{\lambda}_0,i} \ .$$

Notre estimation de $\lambda$ est donc mise à jour en utilisant ce gradient, typiquement en posant :

$$\hat{\lambda}_1 = \hat{\lambda}_0 + \gamma_0 \widehat{\nabla}\mathcal{L}_0^{\hat{\lambda}_0} \ ,$$

pour un certain pas de gradient $\gamma_0$.

**Approximation récursive de $H_t^{\hat{\lambda}_t}$ and $G_t^{\hat{\lambda}_t}$.** Au temps $t$, on simule un $N$-échantillon

$$\{\xi_t^i\}_{1\leq i\leq N} \overset{i.i.d}{\sim} q_t^{\hat{\lambda}_t} \ .$$

$H_t^{\hat{\lambda}_t}(\xi_t^i)$ and $G_t^{\hat{\lambda}_t}(\xi_t^i)$ sont estimés respectivement par

$$\hat{H}_t^{\hat{\lambda}_t,i} = \sum_{j=1}^{N} \bar{w}_{t-1|t}^{\hat{\lambda}_t,i,j} \left( \hat{H}_{t-1}^{\hat{\lambda}_{t-1},j} + \tilde{f}_t^{\hat{\lambda}_t}(\xi_{t-1}^j, \xi_t^i) \right) \ , \tag{10}$$

$$\begin{aligned} \hat{G}_t^{\hat{\lambda}_t,i} = \sum_{i=1}^{N} \bar{w}_{t-1|t}^{\hat{\lambda}_t,i,j} \Big\{ \hat{G}_{t-1}^{\hat{\lambda}_{t-1},j} + \nabla \log q_{t-1|t}^{\lambda_t}(\xi_t^i, \xi_{t-1}^j) \times \\ \left( \hat{H}_{t-1}^{\lambda_{t-1},j} + \tilde{f}_t^{\hat{\lambda}_t}(\xi_{t-1}^j, \xi_t^i) \right) \Big\} \ , \end{aligned} \tag{11}$$

où

$$\bar{w}_{t-1|t}^{\hat{\lambda}_t,i,j} = \frac{q_{t-1|t}^{\hat{\lambda}_t}(\xi_t^i, \xi_{t-1}^j)/q_{t-1}^{\hat{\lambda}_{t-1}}(\xi_{t-1}^j)}{\sum_{k=1}^{N} q_{t-1|t}^{\hat{\lambda}_t}(\xi_t^i, \xi_{t-1}^k)/q_{t-1}^{\hat{\lambda}_{t-1}}(\xi_{t-1}^k)} \ . \tag{12}$$

L'approximation du gradient est alors donnée par l'équation

$$\widehat{\nabla}\mathcal{L}_t^{\hat{\lambda}_t} = \frac{1}{N}\sum_{i=1}^{N} \nabla \log q_t^{\hat{\lambda}_t}(\xi_t^i) \cdot \hat{H}_t^{\hat{\lambda}_t,i} + \hat{G}_t^{\hat{\lambda}_t,i} \ ,$$

qui peut être utilisée pour obtenir $\hat{\lambda}_{t+1}$.

**Remarque sur l'algorithme** Les équations (10) et (11) sont des estimateurs des equations (8) and (9) par *self-normalized importance sampling* (SNIS), dont l'expression des poids auto-normalisés (partagés par les deux estimateurs) est donnée par (12). On notera ici que l'on ne peut pas obtenir des approximations Monte Carlo directes de (8)-(9) en simulant des échantillons selon $q_{t-1|t}^{\lambda}(\xi_t^i, \cdot)$, car les fonctions $H_{t-1}^{\hat{\lambda}_{t-1}}$ and $G_{t-1}^{\hat{\lambda}_{t-1}}$ n'auraient pas été approchées sur ces échantillons. L'*importance sampling* est donc indispensable pour la mise à jour récursive de l'approximation du gradient. On sait cependant que les performances d'un tel estimateur dépendent fortement du lien entre la loi de proposition

3. Á l'étape $t = 0$, il s'agit d'une évaluation exacte, il s'agira ensuite d'approximations

et la loi cible. La Section 4 propose une mise en oeuvre efficace pour relier la la loi de proposition $q_{t-1}^{\hat{\lambda}_{t-1}}$ à la loi cible $q_{t-1|t}^{\hat{\lambda}_t}$. L'auto-normalisation dans (12) est motivée par des considérations pratiques, car elle réduit la variance de l'estimateur dans nos simulations. Cette réduction de la variance vient cependant au prix de l'ajout d'un biais.

Une autre source de biais dans notre approximation est la présence dans les termes de droite de (10)-(11) de quantités ayant été calculées sous le paramètre $\hat{\lambda}_{t-1}$ pour approcher des quantités sous le paramètre $\hat{\lambda}_t$. Ces approximations, couramment effectuées dans des contextes de maximum de vraisemblance récursif (voir [Tadić and Doucet, 2020] par exemple), sont indispensables pour la mise en place pratique de l'apprentissage en ligne. Nos expériences numériques montrent que ces approximations, qui rendraient l'étude théorique plus complexe, conduisent toujours à de bons résultats dans la pratique.

Le lecteur familier des méthodes de Monte Carlo séquentiel pourra voir des similarités avec l'algorithme PaRIS proposé dans [Olsson et al., 2017] et étendu dans [Gloaguen et al., 2022]. Bien qu'inspiré de ces travaux, notre algorithme présente la différence majeure de ne fonctionner qu'avec des échantillons i.i.d., et selon une loi variationelle que l'on choisit (les tirages sont donc directs). Cet aspect évite le problème de la dégénérescence des poids des méthodes de Monte Carlo séquentiel. Dans la pratique, nous nous servons d'astuces de rééchantillonnage des deux articles cités pour éviter d'avoir à calculer la constante de normalisation dans (12), ce qui entrainerait une complexité en $N^2$. Les détails sont disponibles dans [Chagneux et al., 2024b].

# 4 Détails d'implémentation

**Définitions de noyaux *backward* à partir de noyaux *forward*** L'équation (12) suggère que la performance de l'algorithme proposé dépend fortement de la définition des distributions variationnelles et du lien entre $q_{t-1}^{\hat{\lambda}_{t-1}}$ et $q_{t-1|t}^{\hat{\lambda}_t}$. Nous introduisons donc une structure supplémentaire dans la famille variationnelle donnée par (3), en utilisant des fonctions de potentiel $\psi_t^\lambda : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$ pour relier ces distributions. Les fonctions de potentiel $(q_t^\lambda)_{t\geq 0}$ relient explicitement les distributions $(q_t^\lambda)_{t\geq 0}$ aux noyaux rétrospectifs $(q_{t-1|t}^\lambda)_{t\geq 0}$. Plus précisément, nous imposons que, pour tout $t \geq 1$,

$$q_{t-1|t}^{\hat{\lambda}_t}(x_t, x_{t-1}) \propto q_{t-1}^{\hat{\lambda}_{t-1}}(x_{t-1})\psi_t^{\hat{\lambda}_t}(x_{t-1}, x_t) \; . \tag{13}$$

Les fonctions $\psi_t^\lambda$ peuvent être rendues arbitrairement complexes, de sorte que, pour tout $t$, le noyau variationnel *backward* $q_{t-1|t}^{\lambda_t}$ a une dépendance en $x_t$ arbitrairement complexe.

**Paramétrisation des distributions variationelles** En pratique, l'utilisation de la décomposition (13) dans le cadre en ligne nécessite la paramétrisation explicite, pour tout $t \geq 0$, d'une distribution $q_t^\lambda$ et d'un potentiel $\psi_t^\lambda$ (pas nécessairement normalisé) qui dépendent tous deux des observations jusqu'à l'instant $t$ (au maximum). Dans un souci de temps de calcul, chaque $q_t^\lambda$ est choisi comme une densité paramétrique au sein de la famille exponentielle. Cette famille est notée $\mathsf{P} = \{P_\eta\}_{\eta\in\mathcal{E}}$ où $\eta$ est le paramètre naturel correspondant et $\mathcal{E}$, l'espace des paramètres de cette famille (typiquement la famille des distributions gaussiennes définies sur $\mathbb{R}^{d_x}$). On note $\eta_t^\lambda$ le paramètre de $q_t^\lambda$. Pour garantir

que les distributions $q^\lambda_{t-1|t}$ appartiennent à la même famille, nous imposons que

$$\psi^\lambda_t(x_{t-1}, x_t) = \exp\left(\tilde{\eta}^\lambda_t(x_t) \cdot T(x_{t-1})\right) ,$$

où $\tilde{\eta}^\lambda_t(x_t) = \mathsf{MLP}^\lambda(x_t)$ [4] et $T(x_{t-1})$ sont le paramètre naturel et la statistique suffisante pour la famille $\mathsf{P}$. Ainsi, grâce à (13), $q^\lambda_{t-1|t}(x_t, \cdot)$ sera une densité dans $\mathsf{P}$ avec pour paramètre naturel $\eta^\lambda_{t-1|t} = \eta^\lambda_{t-1} + \tilde{\eta}^\lambda_t$. Dans ce cadre pratique, les noyaux $q^\lambda_{t-1|t}$ peuvent avoir des dépendances arbitrairement complexes sur $x_t$ tandis que leur densité est obtenue explicitement à partir des potentiels. Cela élimine la nécessité de calculer des constantes de normalisation (requises, par exemple, lors du calcul de (11)), tout en évitant la réduction de ces noyaux à des familles trop simples (par exemple, les noyaux linéaires gaussiens). Pour les paramètres de $q^\lambda_t$, il existe deux approches principales :

— *Approches amorties* où les paramètres de $q^\lambda_t$ sont mis à jour à l'aide d'une fonction paramétrée à chaque instant $t$. Cela peut se faire à l'aide de quantités intermédiaires $a_t \in \mathsf{A}$ (où $\mathsf{A}$ est un espace défini par l'utilisateur), telles que $a_t = \mathsf{MLP}^\lambda(a_{t-1}, y_t)$, et $\eta^\lambda_t = \mathsf{MLP}^\lambda(a_t)$. L'initialisation est effectuée à l'aide d'un paramètre aléatoire $a_{-1}$, qui peut être fixe ou appris. Les schémas amortis sont efficaces d'un point de vue numérique (puisque les connaissances des prédictions précédentes sont utilisées pour produire les paramètres actuels), mais nécessitent la définition manuelle de fonctions complexes. Les récursions peuvent être analytiques (et ne pas reposer sur un MLP), par exemple lorsque $q^\lambda_{0:t}$ est la distribution de lissage d'une gaussienne linéaire, ou lorsque la conjugaison est davantage exploitée pour mettre à jour les paramètres $(\eta^\lambda_t)_{t\geq 0}$ (voir l'annexe **??**). Quoi qu'il en soit, le nombre de paramètres devient indépendant de $t$, cependant la rétropropagation des gradients va mécaniquement croître linéairement avec $t$. Pour éviter cela, une solution consiste à tronquer la rétropropagation, c'est-à-dire à supposer que $(a^\lambda_s)_{s\leq t-\Delta}$ est indépendant de $\lambda$ pour un certain $\Delta$.

— *Approches non amorties* où chaque $q^\lambda_t$ et $\psi^\lambda_t$ ont leurs propres paramètres $\eta_t$ et $\tilde{\eta}_t$, sans rapport avec ceux du temps $t-1$. Dans ce cas, le vecteur optimisé $\lambda$ contient les paramètres $(\eta_t)_{t\geq 0}$, et le nombre de paramètres croît alors linéairement avec $t$. Ce schéma modifie l'équation (9) (voir [Chagneux et al., 2024b] pour plus de détails).

**Réduction de variance** Les équations (7) et (9) nécessite l'approximation d'espérance selon une fonction de score, i.e. des espérances de la forme $\mathbb{E}_{q^\lambda}\left[\nabla \log q^\lambda(X) \cdot f(X)\right]$, pour une certaine distribution $q^\lambda(X)$. Comme discuté dans [Mohamed et al., 2020], l'estimateur Monte Carlo direct est suvent sujet à une grande variance, et il est souhaitable, pour un bonne performance pratique, d'essayer de réduire la variance en introduisant une variable de contrôle. En remarquant que $\mathbb{E}_{q^\lambda}\left[\nabla \log q^\lambda(X)\right] = 0$, notre espérance cible est donc égale à $\mathbb{E}_{q^\lambda}[\nabla \log q^\lambda(X)(f(X) - \mathbb{E}_{q^\lambda}[f(X)])]$. Or, des estimations Monte Carlo de $\mathbb{E}_{q^\lambda_t}[H^\lambda_t]$ et $\mathbb{E}_{q^\lambda_{t-1|t}}[H^\lambda_{t-1}(X_{t-1}) + \tilde{f}^\lambda_t(X_{t-1}, x_t)]$ sont calculés dans l'algorithme de la Section 3, à savoir $N^{-1}\sum_{i=1}^N \hat{H}^{\lambda,i}_t$ et $\{\hat{H}^{\lambda,i}_t\}_{1\leq i\leq N}$. Notre méthodologie amène donc directement à une réduction de variance efficace.

4. $\mathsf{MLP}$ est une notation pour désigner un *multi-layer perceptron*

# Références


[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference : A review for statisticians. *Journal of the American Statistical Association*, 112(518) :859–877.

[Campbell et al., 2021] Campbell, A., Shi, Y., Rainforth, T., and Doucet, A. (2021). Online variational filtering and parameter learning. *Advances in Neural Information Processing Systems*, 34.

[Chagneux et al., 2024a] Chagneux, M., Gassiat, É., Gloaguen, P., and Le Corff, S. (2024a). Additive smoothing error in backward variational inference for general state-space models. *Journal of Machine Learning Research*.

[Chagneux et al., 2024b] Chagneux, M., Gloaguen, P., Corff, S. L., and Olsson, J. (2024b). Importance sampling for online variational learning.

[Gloaguen et al., 2022] Gloaguen, P., Corff, S. L., and Olsson, J. (2022). A pseudo-marginal sequential Monte Carlo online smoothing algorithm. *Bernoulli*, 28(4) :2606 – 2633.

[Mohamed et al., 2020] Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2020). Monte carlo gradient estimation in machine learning. *The Journal of Machine Learning Research*, 21(1) :5183–5244.

[Olsson et al., 2017] Olsson, J., Westerborn, J., et al. (2017). Efficient particle-based online smoothing in general hidden markov models : the PaRIS algorithm. *Bernoulli*, 23(3) :1951–1996.

[Tadić and Doucet, 2020] Tadić, V. Z. and Doucet, A. (2020). Asymptotic properties of recursive particle maximum likelihood estimation. *IEEE Transactions on Information Theory*, 67(3) :1825–1848.


# A Preuve de la Proposition 1

En partant de la définition de l'ELBO, on a que :

$$
\begin{aligned}
\mathcal{L}_t^\lambda &= \mathbb{E}_{q_{0:t}^\lambda}\left[\log \frac{p_{0:t}(X_{0:t}, Y_{0:t})}{q_{0:t}^\lambda(X_{0:t})}\right] \\
&= \mathbb{E}_{q_{0:t}^\lambda}\left[\sum_{s=0}^{t} \tilde{f}_s^\lambda(X_{s-1}, X_s)\right] - \mathbb{E}_{q_t^\lambda}\left[\log q_t^\lambda(X_t)\right] \\
&= \int \left\{\sum_{s=0}^{t} \tilde{f}_s^\lambda(x_{s-1}, x_s)\right\} q_t^\lambda(x_t) \prod_{s=1}^{t} q_{s-1|s}^\lambda(x_s, x_{s-1}) \mathrm{d}x_{1:t} - \mathbb{E}_{q_t^\lambda}\left[\log q_t^\lambda(X_t)\right] \\
&= \mathbb{E}_{q_t^\lambda}\left[H_t^\lambda(X_t)\right] - \mathbb{E}_{q_t^\lambda}\left[\log q_t^\lambda(X_t)\right] \ ,
\end{aligned}
$$

où

$$
H_t^\lambda(x_t) := \int \left\{\sum_{s=0}^{t} \tilde{f}_s^\lambda(x_{s-1}, x_s)\right\} \prod_{s=1}^{t} q_{s-1|s}^\lambda(x_s, x_{s-1}) \mathrm{d}x_{1:t-1} = \mathbb{E}_{q_{0:(t-1)|t}^\lambda}\left[f_{0:t}^\lambda(X_{0:t-1}, x_t)\right] \ ,
$$

où

$$
q_{0:(t-1)|t}^\lambda(x_{0:t-1}, x_t) = \prod_{s=1}^{t} q_{s-1|s}^\lambda(x_s, x_{s-1})
$$

est la densité de $X_{0:t-1}$ conditonnellement à $X_t = x_t$. On remarque ensuite que

$$\begin{aligned}
H_t^\lambda(x_t) &= \int \left( \int \left\{ \sum_{s=0}^{t} \tilde{f}_s^\lambda(x_{s-1}, x_s) \right\} \prod_{s=1}^{t-1} q_{s-1|s}^\lambda(x_s, x_{s-1}) \mathrm{d}x_{1:t-2} \right) q_{t-1|t}^\lambda(x_t, x_{t-1}) \mathrm{d}x_{t-1} \\
&= \int \left( H_{t-1}^\lambda(x_{t-1}) + \tilde{f}_t^\lambda(x_{t-1}, x_t) \right) q_{t-1|t}^\lambda(x_t, x_{t-1}) \mathrm{d}x_{t-1} \\
&= \mathbb{E}_{q_{t-1|t}^\lambda} \left[ H_{t-1}^\lambda(X_{t-1}) + \tilde{f}_t^\lambda(X_{t-1}, X_t) \right] .
\end{aligned}$$

Les égalités (6) et (8) sont donc établies. Considérons désormais le gradient :

$$\begin{aligned}
\nabla \mathcal{L}_t^\lambda =& \nabla \mathbb{E}_{q_t^\lambda} \left[ H_t^\lambda(X_t) \right] - \nabla \mathbb{E}_{q_t^\lambda} \left[ \log q_t^\lambda(X_t) \right] \\
=& \mathbb{E}_{q_t^\lambda} \left[ \nabla H_t^\lambda(X_t) \right] - \mathbb{E}_{q_t^\lambda} \left[ \nabla \overbrace{\log q_t^\lambda(X_t)}^{=0} \right] + \\
& \int \left( H_t^\lambda(x_t) - \log q_t^\lambda(x_t) \right) \nabla \left( \log q_t^\lambda(x_t) \right) q_t^\lambda(x_t) \mathrm{d}x_t \\
=& \mathbb{E}_{q_t^\lambda} \left[ \nabla H_t^\lambda(X_t) + \left( H_t^\lambda(x_t) - \log q_t^\lambda(x_t) \right) \times \nabla \log q_t^\lambda(x_t) \right] ,
\end{aligned}$$

où $\mathbb{E}_{q_t^\lambda} \left[ \nabla \log q_t^\lambda(X_t) \right] = 0$ car $\mathbb{E}_{q_t^\lambda} \left[ \nabla \log q_t^\lambda(X_t) \right] = \int \nabla q_t^\lambda(x_t) \mathrm{d}x_t = 0$. En notant $G_t^\lambda(x_t) = \nabla H_t^\lambda(x_t)$, on retrouve don l'égalité (7). On a de plus que :

$$\begin{aligned}
G_t^\lambda(x_t) &= \nabla \mathbb{E}_{q_{0:(t-1)|t}^\lambda} \left[ f_{0:t}^\lambda(X_{0:t-1}, x_t) \right] \\
&= \mathbb{E}_{q_{0:(t-1)|t}^\lambda} \left[ \left\{ \nabla \log q_{0:(t-1)|t}^\lambda \times f_{0:t}^\lambda \right\} (X_{0:t-1}, x_t) \right] + \mathbb{E}_{q_{0:(t-1)|t}^\lambda} \left[ \nabla f_{0:t}^\lambda(X_{0:t-1}, x_t) \right] .
\end{aligned}$$

Or ici, on remarque [5] que, par definition de $f_{0:t}^\lambda$ :

$$\nabla f_{0:t}^\lambda(X_{0:t-1}, x_t) = -\nabla \log q_{0:(t-1)|t}^\lambda(X_{0:t-1}, x_t),$$

donc, l'espérance de ce terme est égale à 0 et :

$$G_t^\lambda(x_t) = \mathbb{E}_{q_{0:(t-1)|t}^\lambda} \left[ \left\{ \nabla \log q_{0:(t-1)|t}^\lambda \times f_{0:t}^\lambda \right\} (X_{0:t-1}, x_t) \right] .$$

En développant la fonction dans l'espérance, on aboutit à :

$$\begin{aligned}
G_t^\lambda(x_t) =& \mathbb{E}_{q_{t-1|t}^\lambda} \left[ G_{t-1}^\lambda(X_{t-1}) \right] \\
& + \mathbb{E}_{q_{t-1|t}^\lambda} \left[ \nabla \log q_{t-1|t}^\lambda(X_{t-1}, x_t) \left( \mathbb{E}_{q_{0:(t-2)|t-1}^\lambda} \left[ f_{0:t-1}^\lambda(X_{0:t-1}) \right] + \tilde{f}_t^\lambda(X_{t-1}, x_t) \right) \right] \\
& + \mathbb{E}_{q_{t-1|t}^\lambda} \left[ \tilde{f}_t^\lambda(X_{t-1}, x_t) \times \mathbb{E}_{q_{0:(t-2)|t-1}^\lambda} \left[ \nabla \log q_{0:(t-2)|t-1}^\lambda(X_{0:t-1}) \right] \right] .
\end{aligned}$$

Dans l'espérance de la deuxième ligne, on reconnait $H_{t-1}^\lambda$ et la troisème ligne est encore égale à 0. Ce qui montre (9).

5. Ceci est spécifique à l'ELBO, car le numérateur de dépend pas de $\lambda$

# Non-asymptotic confidence intervals for importance sampling estimators of quantiles

Baalu Ketema[1,2,*], Roman Sueur[1], Nicolas Bousquet[1,3,4], Bertrand Iooss[1,2,3], Fabrice Gamboa[2], Francesco Costantino[2]

[1] *EDF R&D, 6 Quai Watier, 78400 Chatou, France*
[2] *Institut de Mathématiques de Toulouse, 31062, Toulouse, France*
[3] *SINCLAIR AI Laboratory, Saclay, France*
[4] *Sorbonne Université, LPSM, 4 place Jussieu, Paris, France*

[*] *Corresponding author - baalu-belay.ketema@edf.fr*

**Résumé.** La construction d'intervalle de confiance (asymptotiques ou non-asymptotiques) est une étape cruciale pour comprendre la qualité de l'estimation d'une quantité d'intérêt bâtie sur une distribution. Dans cette présentation, nous estimons un quantile $q_\alpha$ d'une variable aléatoire réelle $Y \sim \mu$ dans le cas où seul un échantillon d'une autre distribution $\mu_0$ est disponible et où $\mu_0$ domine $\mu$. La méthode d'estimation utilisée est l'échantillonnage préférentiel. Un TCL est connu pour l'estimateur du quantile mais la variance asymptotique dépend du quantile $q_\alpha$ de $\mu$, l'inconnu, et de sa fonction de répartition $F_\mu$. Nous levons ce verrou en construisant un intervalle de confiance non-asymptotique pour $q_\alpha$ qui peut être utile lorsque l'on ne dispose que d'un échantillon de taille limitée.

**Mots-clés.** Echantillonnage préférentiel, estimation de quantile, inégalité de concentration, intervalles de confiance non-asymptotique.

**Abstract.** Building a confidence region (asymptotic or non-asymptotic) is crucial in understanding the quality of point estimators of a distribution. In this presentation, we estimate a quantile $q_\alpha$ of a real random variable $Y \sim \mu$ in the case where only a sample from another distribution $\mu_0$ is available and where $\mu_0$ dominates $\mu$. This estimation procedure is known as importance sampling. A CLT is proved for the quantile estimator but the asymptotic variance depends on the quantile $q_\alpha$ of $\mu$, the unknown, and on its cumulative distribution function $F_\mu$. We lift this barrier by building a non-asymptotic confidence interval for $q_\alpha$ which can be useful when only a limited sample size is available.

**Keywords.** Importance sampling, quantile estimation, concentration inequality, non-asymptotic confidence intervals.

## 1 Introduction

In many industrial contexts, quantities of interest (QoI) are defined from real variables $Y \sim \mu$ considered as random with underlying distribution $\mu$, that represent the behavior of a component or system. For instance $Y$ is the output of a code that computes the level of a river [7] or the cladding temperature in a nuclear vessel after an accident [4]. Typical QoIs are the quantile $q_\alpha(Y)$, the superquantile $Q_\alpha(Y)$ [6] or a probability $p_T = \mathbb{P}(Y > T)$ given some threshold $T$. Usually these QoIs cannot be computed explicitly if $\mu$ cannot be easily

handled (ie., not in closed form). Hence statistical estimation is required to approximate these quantities. In addition, confidence regions are usually built to understand how far is the estimator to the real value.

The standard estimation method uses a Monte Carlo simulation to approximate the cumulative distribution function (cdf) of $Y$ denoted $F_\mu$: for $Y_1, ..., Y_N$ an iid sample from $\mu$, we have that

$$\widehat{F}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$$

is a non biased estimator of the cdf of $Y$ for all $t$ in $\mathbb{R}$. It converges a.s. and uniformly in $t$ to $F_\mu$. It also verifies a functional central limit theorem ([9], Chapter 19). We can therefore build plug-in estimators for the quantile $q_\alpha(Y)$, the superquantile $Q_\alpha(Y)$ and the probability threshold $p_T$:

$$\begin{aligned}
\widehat{q}_\alpha &:= \inf\{t \in \mathbb{R} \mid \widehat{F}(t) \geq \alpha\}, \\
\widehat{Q}_\alpha &:= \frac{1}{N(1-\alpha)}\sum_{i=1}^{N} Y_i \mathbf{1}_{Y_i \geq \widehat{q}_\alpha}, \\
\widehat{p}_T &:= 1 - \widehat{F}(T).
\end{aligned}$$

The associated presentation only focuses on the quantile $q_\alpha(Y)$. From [9], Chapter 21, the estimator $\widehat{q}_\alpha$ satisfies a central limit theorem: assuming that $F_\mu$ is differentiable at point $q_\alpha := q_\alpha(Y)$ with $F'_\mu(q_\alpha) > 0$ then we have that

$$\sqrt{N}\big(\widehat{q}_\alpha - q_\alpha\big) \to \mathcal{N}(0, \sigma^2_\infty),$$

where

$$\sigma^2_\infty = \frac{\alpha(1-\alpha)}{F'_\mu(q_\alpha)^2}.$$

The asymptotic variance depends on $q_\alpha = q_\alpha(Y)$, the unknown QoI, as well as on $F_\mu$ and therefore this result cannot be used directly to construct asymptotic confidence intervals. In addition, obtaining a large sample of $Y$, a requirement for asymptotic confidence intervals, can be very time-consuming (for instance $Y$ can be the output of a costly industrial computer code ie., $Y = G(X)$ and a sample of $Y$ is obtained by evaluating $G$ on a sample of $X$). Consequently, non-asymptotic confidence intervals are more appropriate in this case, as they provide information on the concentration of the estimator around the true value as a function of the sample size $N$.

The following paper is organized as follows. Section 2 discusses how non-asymptotic confidence intervals can be built for the standard quantile estimator using a uniform concentration inequality on the empirical cdf. Section 3 explains the importance sampling method for the quantile estimator and states a CLT for the latter. Section 4 states the main result of the paper ie., a method for building non-asymptotic confidence intervals for importance sampling estimators of quantiles. And lastly, Section 5 discusses the limitations of this method.

## 2 Concentration inequality for the quantile estimator

The Dvoretzky–Kiefer–Wolfowitz (DKW) theorem [3, 8] shows that the estimator $\widehat{F}$ of $F_\mu$ verifies the following concentration inequality:

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |\widehat{F}(t) - F_\mu(t)| > \eta\right) \leq 2e^{-2N\eta^2},$$

for all $\eta > 0$. This concentration inequality can be used to obtain a non-asymptotic confidence interval for the quantile estimator $\widehat{q}_\alpha$: for all $\eta > 0$ small enough and $\alpha \in (0,1)$ fixed

$$\mathbb{P}\Big(\widehat{q}_{\alpha-\eta} \leq q_\alpha(Y) \leq \widehat{q}_{\alpha+\eta}\Big) \geq 1 - 2e^{-2N\eta^2}, \tag{1}$$

since DKW implies that with probability at least $1 - 2e^{-2N\eta^2}$, the functional inequality

$$\widehat{F} - \eta \leq F_\mu \leq \widehat{F} + \eta \tag{2}$$

is verified. And since $\widehat{F}$ and $F_\mu$ are non-decreasing functions, taking the generalized inverse in (2) gives for all $\alpha \in (0,1)$

$$(\widehat{F} - \eta)^{\langle -1\rangle}(\alpha) \geq F_\mu^{\langle -1\rangle}(\alpha) = q_\alpha(Y) \geq (\widehat{F} + \eta)^{\langle -1\rangle}(\alpha) \tag{3}$$

with probability at least $1-2e^{-2N\eta^2}$, where $H^{\langle -1\rangle}$ is the generalized inverse of a non-decreasing function $H$ defined as $H^{\langle -1\rangle}(\alpha) := \inf\{t \in \mathbb{R} \mid H(t) \geq \alpha\}$. The non-asymptotic confidence interval's quality heavily depends on the sample size $N$ of $Y$, see Table 1, as well as on the order of the quantile $\alpha$ ie., for $\alpha$ close to 0 or 1 a very large sample will be required to accurately approximate the quantile.

| Sample size $N$ | $N = 10^4$ | $N = 10^5$ | $N = 2 \times 10^6$ |
|---|---|---|---|
| Confidence level $\geq 0.75$ | [1.52, 1.71] | [1.62, 1.68] | [1.635, 1.649] |
| Confidence level $\geq 0.95$ | [1.50, 1.76] | [1.61, 1.69] | [1.632, 1.651] |
| Confidence level $\geq 0.99$ | [1.48, 1.79] | [1.60, 1.70] | [1.631, 1.654] |

Table 1: Confidence intervals (1) on the 0.95-quantile of $\mathcal{N}(0,1)$ in terms of the sample size and a fixed confidence level. The value of this quantile is approximately 1.6448.

## 3 Importance sampling estimation procedure

Assume we do not have access to a sample of $\mu$ but rather a sample from another distribution $\mu_0$ on $\mathbb{R}$ which dominates $\mu_0$ (ie., $\mu << \mu_0$ meaning that $\mu$ admits a density on $\mathbb{R}$ with respect to (wrt) $\mu_0$). Denote $L := \frac{d\mu}{d\mu_0}$ the Radon-Nikodym derivative (also called the likelihood ratio). We would like to build estimators of $q_\alpha$, a quantile of $\mu$, as well as confidence intervals using an iid sample $Y_1, ..., Y_N$ of $\mu_0$. To do so we can use the importance sampling (IS) method

$$\widehat{F}(t) := \frac{1}{N}\sum_{i=1}^{N} L(Y_i)\mathbf{1}_{Y_i \leq t},$$

which is the standard unbiased Monte Carlo estimator of $F_\mu(t)$. But $\widehat{F}$ is not the cdf of a discrete measure on $\mathbb{R}$ since the weights $\frac{L(Y_i)}{N}$ do not add up to one. Hence we favor instead the following biased estimator

$$\widehat{F}_{\text{is}}(t) := \frac{1}{\sum_{i=1}^N L(Y_i)} \sum_{i=1}^N L(Y_i) \mathbf{1}_{Y_i \leq t},$$

which also converges pointwise a.s. to $F_\mu$. It allows us to build an estimator of the quantile of $\mu$ by plug-in:

$$\widehat{q}_\alpha^{\,\text{is}} := \inf\{t \in \mathbb{R} \mid \widehat{F}_{\text{is}}(t) \geq \alpha\}.$$

The asymptotic properties of this estimator are already studied in [4, 5], who showed that if (a) $L$ is cube-integrable wrt $\mu_0$; (b) $F_\mu$ is differentiable at $q_\alpha := q_\alpha(Y)$ for $Y \sim \mu$; (c) $F'_\mu(q_\alpha) > 0$, then

$$\sqrt{N}\big(\widehat{q}_\alpha^{\,\text{is}} - q_\alpha\big) \underset{N\to\infty}{\longrightarrow} \mathcal{N}(0, \sigma_\infty^2),$$

where

$$\sigma_\infty^2 = \frac{\mathbb{E}_{Y\sim\mu_0}\big[L(Y)^2(\mathbf{1}_{Y\leq q_\alpha} - \alpha)^2\big]}{F'_\mu(q_\alpha)^2}.$$

This asymptotic variance depends again on the unknown QoI $q = q_\alpha(Y)$. Hence this result cannot be directly used to build asymptotic confidence intervals.

# 4 Non-asymptotic confidence intervals for the IS quantile estimator

Now our goal is to build non-asymptotic confidence intervals for the quantile estimator $\widehat{q}_\alpha^{\,\text{is}}$. Assume we have a pointwise confidence interval for $F_\mu$ around $\widehat{F}_{\text{is}}$ ie., for each fixed $t$ in $\mathbb{R}$, for all $N \in \mathbb{N}^*$ there exist $\varepsilon_N > 0$ and $\lambda_{t,N}^- < \lambda_{t,N}^+$ such that

$$\mathbb{P}\big(F_\mu(t) \in [\lambda_{t,N}^-, \lambda_{t,N}^+]\big) \geq 1 - \varepsilon_N,$$

where $\varepsilon_N \to 0$ as $N \to \infty$, $t \to \lambda_{t,N}^\pm$ are non-decreasing random functions and $\lambda_{t,N}^-, \lambda_{t,N}^+ \underset{N\to\infty}{\longrightarrow} F_\mu(t)$ a.s.. Then we can prove the following result.

**Theorem.** *Under the previous assumptions, by choosing a decreasing sequence $(a_N)_{N\geq 1}$ such that $a_N \leq 1$ and $\lim_{N\to\infty} a_N = 0$ we have:*

$$\mathbb{P}\big(q_\alpha^- \leq q_\alpha(Y) \leq q_\alpha^+\big) \geq 1 - \left\lfloor \frac{1}{a_N} - 1 \right\rfloor \varepsilon_N,$$

*where $q_\alpha^\mp$ are the generalized inverse of the functions $t \to \lambda_{t,N}^\pm \pm a_N$ evaluated at $\alpha \in (0,1)$. They are random variables taking values in the original sample $\{Y_1, ..., Y_N\} \cup \{\pm\infty\}$ of $\mu_0$.*

Note that this theorem does not guaranty that the quantity $1 - \left\lfloor \frac{1}{a_N} - 1 \right\rfloor \varepsilon_N$ is positive. This depends on the sequence $a_N$ which has to be cleverly chosen. The proof is inspired from [2] and is based on the following two ingredients:

(i) we can convert the pointwise confidence interval into a uniform one by slightly enlarging it by doing $\lfloor \frac{1}{a_N} - 1 \rfloor$ union bounds:

$$\mathbb{P}\left(\cap_{t\in\mathbb{R}} \left\{F_\mu(t) \in [\lambda^-_{t,N} - a_N, \lambda^+_{t,N} + a_N]\right\}\right) \geq 1 - \left\lfloor \frac{1}{a_N} - 1 \right\rfloor \varepsilon_N,$$

(ii) we then transform the inequality

$$\lambda^-_{t,N} - a_N \leq F_\mu(t) \leq \lambda^+_{t,N} + a_N, \quad \forall t \in \mathbb{R},$$

into an inequality on the quantile $q_\alpha(Y)$ by means of taking the generalized inverse wrt $t$ as in (3).

Now, in order to obtain the $\lambda^\pm_{t,N}$ for the pointwise confidence interval of $F_\mu(t)$, for all $t$, we can apply Theorem 2 of [1]: for all $s > 0$ take $N = e^{\mathcal{D}(\mu|\mu_0)+s}$, then for all $t \in \mathbb{R}$ we have

$$\mathbb{P}\left(\left|\widehat{F}_{\text{is}}(t) - F_\mu(t)\right| \geq \frac{2\varepsilon_s\sqrt{F_\mu(t)}}{1-\varepsilon_s}\right) \leq 2\varepsilon_s, \tag{4}$$

where $\mathcal{D}$ is the Kulback-Leibler divergence and $\varepsilon_s$ is given by

$$\varepsilon_s = \left(e^{-s/4} + \sqrt{\mathbb{P}\big(\log L(Y) > \mathcal{D}(\mu|\mu_0) + s/2\big)}\right)^{1/2},$$

where $\log L(Y)$ is the log-likelihood ratio of $\mu$ and $\mu_0$ evaluated at $Y \sim \mu_0$.

The concentration inequality (4) can be equivalently rewritten as

$$\mathbb{P}\Big(F_\mu(t) \in [\lambda^-_{t,s}, \lambda^+_{t,s}]\Big) \geq 1 - 2\varepsilon_s,$$

where

$$\lambda^\pm_{t,s} := \frac{2\widehat{F}_{\text{is}}(t) + \eta_s^2 \pm \eta_s\sqrt{4\widehat{F}_{\text{is}}(t) + \eta_s^2}}{2},$$

and $\eta_s := \frac{2\varepsilon_s}{1-\varepsilon_s}$. Indeed $\lambda^\pm_{t,s}$ verify the necessary assumptions of Theorem 4.

This method can also be used when we have a parametric family $\mathcal{P} := \{\mu_\theta : \theta \in \Theta\}$ and $\mu, \mu_0 \in \mathcal{P}$ and we have an iid sample $Y_1, ..., Y_N$ wrt $\mu_{\theta_0} := \mu_0$, and we want to build an estimator of a quantile on $\mu_\theta$ and a corresponding confidence interval for all $\theta \in \Theta$.

# 5 Limitations

The quality of the confidence interval for the quantile $q_\alpha(Y)$ depends on the quality of the initial pointwise confidence interval on $F_\mu$, built from a sample of $\mu_0$. Indeed, the authors of [1] specifically mention that no efforts were made to improve the concentration inequality (4) so the pointwise confidence interval for $F_\mu(t)$ is not necessarily good. In addition, choosing the sequence $(a_N)_N$ is not obvious and requires a compromise between having a small uniform confidence interval around $F_\mu$ and a high confidence level. Moreover, the confidence interval obtained for the quantile is actually uniform in $\alpha \in (0,1)$ ie.,

$$\mathbb{P}\Big( \cap_{\alpha\in(0,1)} \left\{ q_\alpha^- \leq q_\alpha(Y) \leq q_\alpha^+ \right\} \Big) \geq 1 - \left\lfloor \frac{1}{a_N} - 1 \right\rfloor \varepsilon_N.$$

This is because we inverted a uniform bound in $t$ on the cdf. This means that if we want a confidence interval for a specific $\alpha$, for instance $\alpha = 0.95$, then the confidence level $1 - \left\lfloor \frac{1}{a_N} - 1 \right\rfloor \varepsilon_N$ might be too conservative since

$$\begin{aligned} \mathbb{P}\Big( q_{0.95}^- \leq q_{0.95}(Y) \leq q_{0.95}^+ \Big) &\geq \mathbb{P}\Big( \cap_{\alpha\in(0,1)} \left\{ q_\alpha^- \leq q_\alpha(Y) \leq q_\alpha^+ \right\} \Big) \\ &\geq 1 - \left\lfloor \frac{1}{a_N} - 1 \right\rfloor \varepsilon_N. \end{aligned} \tag{5}$$

Therefore more work is needed to understand how much is lost at (5).

# References


[1] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

[2] M. Ducoffe, S. Gerchinovitz, and J. S. Gupta. A high probability safety guarantee for shifted neural network surrogates. In *SafeAI@ AAAI*, pages 74–82, 2020.

[3] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[4] C. Gauchy, J. Stenger, R. Sueur, and B. Iooss. An information geometry approach to robustness analysis for the uncertainty quantification of computer codes. *Technometrics*, 64(1):80–91, 2022.

[5] P.W. Glynn. Importance sampling for monte carlo estimation of quantiles. In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, pages 180–185. Citeseer, 1996.

[6] B. Iooss, V. Vergès, and V. Larget. Bepu robustness analysis via perturbed law-based sensitivity indices. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 236(5):855–865, 2022.

[7] P. Lemaître, E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, and B. Iooss. Density modification-based reliability sensitivity analysis. *Journal of Statistical Computation and Simulation*, 85(6):1200–1223, 2015.

[8] P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[9] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

# A GRADIENT APPROXIMATION WITH IMPORTANCE SAMPLING FOR DIMENSION REDUCTION IN NATURAL EXPONENTIAL FAMILIES

Bastien Batardière, [1] Julien Chiquet [1], Joon Kwon [1]& Julien Stoehr [2]

[1] *MIA Paris-Saclay, Paris-Saclay University, AgroParisTech, INRAE, France, {prenom.nom}@inrae.fr*

[2] *Ceremade, Paris-Dauphine University, France, stoehr@ceremade.dauphine.fr*

**Résumé.** Les données de comptage en grande dimension sont difficiles à analyser telles quelles, et les approches fondées sur des modèles statistiques à variable latente restent efficaces et appropriées, tout en préservant l'explicabilité. Nous considérons plus particulièrement ici le cadre de modèles où les données discrètes sont guidées par une variable gaussienne latente décrivant la structure de dépendances des comptages dans un espaces de faible dimension, puis envoyées dans un espace de grande dimension via une distribution de Poisson ou Binomiale. Comme la loi de la variable latente conditionnement aux données reste inconnue, l'inférence variationnelle s'est révélée efficace pour inférer un tel modèle. Cependant, elle ne maximise qu'une borne inférieure de la vraisemblance et les estimateurs correspondant souffrent d'un manque de garanties théoriques. De plus, un grand nombre de paramètres variationnels est nécessaires. Dans ce travail en cours, nous utilisons l'échantillonnage préférentiel pour estimer les gradients de la log-vraisemblance. Nous contrôlons le biais de l'estimateur et nous appuyons sur des théorèmes d'optimisation pour assurer la convergence d'un schéma de gradient stochastique, s'adaptant facilement à un grand nombre d'échantillons.

**Mots-clés.** Données de comptage, optimisation, échantillonage préférentiel, descente de gradient, famille exponentielle naturelle.

**Abstract.** High-dimensional counting data is challenging to analyze as is, and approaches based on latent variable statistical models remain effective and appropriate, while preserving explainability. We particularly consider here the framework of models where discrete data are guided by a latent Gaussian variable describing the dependency structure of counts in a low-dimensional space, and then mapped into a high-dimensional space via a Poisson or Binomial distribution. Since the law of the latent variable conditioned on the data remains unknown, variational inference has proven effective in inferring such a model. However, it only maximizes a lower bound of the likelihood, and the corresponding estimators suffer from a lack of theoretical guarantees. Additionally, a large number of variational parameters are required. In this ongoing work, we use importance sampling to estimate the gradients of the log-likelihood. We control the bias of the estimator and rely on optimization theorems to ensure the convergence of a stochastic gradient scheme, easily adapting to a large number of samples.

**Keywords.** Count data, optimisation, importance sampling, gradient descent, natural exponential family

# 1 Model

Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip}) \in \mathbb{N}^p$ be some high-dimensional observation vectors of counts, for individual $i$ varying in $1 \leq i \leq n$. We consider a model relying on a multivariate centered scaled Gaussian low dimensional latent variable $\mathbf{W}_i \in \mathbb{R}^q$ ($q \ll p$) linked to the observations $\mathbf{Y}_i$ through a linear function $f_{\theta,i}$ ($1 \leq i \leq n$) with $\theta \in \mathbb{R}^d$ a vector of parameters. Conditionally on $\mathbf{Z}_i = f_{\theta,i}(\mathbf{W}_i) \in \mathbb{R}^p$, the distribution of the observations is assumed to belong to the natural exponential family (NEF). Formally,

$$\begin{aligned} \text{latent space} \quad & \mathbf{W}_i \sim^{\text{iid}} \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q), \quad \mathbf{Z}_i = f_{\theta,i}(\boldsymbol{W}_i) = \mathbf{C}\mathbf{W}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{O}_i, \\ \text{observation space} \quad & p_\theta\left(Y_{ij}|Z_{ij}\right) = \exp(Y_{ij}Z_{ij} - A(Z_{ij}) - h(Y_{ij})), \quad 1 \leq j \leq p\,, \end{aligned} \tag{1}$$

where $h$ and $A$ are real-valued functions with $A$ convex and differentiable and $\theta = (\boldsymbol{C}, \boldsymbol{\beta})$ with $\mathbf{C} \in \mathbb{R}^{p\times q}, \mathbf{X}_i \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^{m\times p}, \mathbf{O}_i \in \mathbb{R}^p$. The parameter $\theta \in \mathbb{R}^d$ with $d = p(q+m)$ is not identifiable as multiplying $\boldsymbol{C}$ by an orthogonal matrix leaves the model unchanged and only $(\boldsymbol{\Sigma}, \boldsymbol{\beta})$ with $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^\top$ is identifiable. We consider 2 distributions inside the NEF, namely the Poisson and Binomial distributions. Every function is known and only the parameter $\theta$ is unknown, which is to be estimated from data $(\mathbf{Y}_i)_{1\leq i\leq n}$.

# 2 Estimation

A natural strategy to estimate $\theta$ is by maximizing the log-likelihood function $\ell(\cdot)$ defined as the finite sum of the log-likelihood of each observation:

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i(\theta) = \frac{1}{n}\sum_{i=1}^n \log p_\theta(\mathbf{Y}_i). \tag{2}$$

**Proposition 2.1.** *For all $1 \leq i \leq n$, $\theta \mapsto \ell_i(\theta)$ is $\mathcal{C}^1$.*

## 2.1 Biased Stochastic Gradient Descent

Given $T \geq 1, \eta > 0$ and $\theta^{(0)} \in \mathbb{R}^d$, one can recursively define $\theta^{(t)}$ via Stochastic Gradient Descent:

$$\theta^{(t+1)} = \theta^{(t)} + \eta\widehat{g}^{(t)} \tag{SGD}$$

where, at each iteration $t \geq 0$, $\widehat{g}^{(t)}$ is a (possibly biased) estimator of $\nabla_\theta \ell(\theta^{(t)})$. An estimator is suggested in the next paragraph.

**Constructing the Estimator** Let $N \geq 1$ be a number of Monte-Carlo samples, $\{\pi(\cdot, \theta, i)\}_{\theta\in\mathbb{R}^d}$ a family of positive density on $\mathbb{R}^q$ with $1 \leq i \leq n$. An index $i(t) \sim \text{Unif}\{1, \cdots, n\}$ is sampled and conditionnally on $\theta^{(0)}, \cdots, \theta^{(t)}, i(t)$, $\left(\mathbf{V}_k^{(t)}\right)_{1\leq k\leq N} \overset{\text{iid}}{\sim} \pi^{(t)}$ with $\pi^{(t)} \triangleq \pi(\cdot, \theta^{(t)}, i(t))$ and

$$\widehat{g}^{(t)} \triangleq \sum_{k=1}^N \omega_k^{(t)} \nabla_\theta \log p_{\theta^{(t)}}(\mathbf{Y}_{i(t)}, \mathbf{V}_k^{(t)}), \text{ where } \omega_k^{(t)} = \frac{\rho_k^{(t)}}{\sum_{\ell=1}^N \rho_\ell^{(t)}} \text{ with } \rho_k^{(t)} = \frac{p_{\theta^{(t)}}(\mathbf{Y}_{i(t)}, \mathbf{V}_k^{(t)})}{\pi^{(t)}(\mathbf{V}_k^{(t)})}. \tag{3}$$

We denote $G^{(t)} = \otimes_{k=1}^{N} \pi^{(t)}$ and $\mathbf{V}^{(t)} \triangleq \left(\mathbf{V}_k^{(t)}\right)_{1 \leq k \leq N}$ so that conditionnaly on $\theta^{(0)}, \ldots, \theta^{(t)}$ and $i(t)$, $V^{(t)} \sim G^{(t)}$. The resulting algorithm when Eq. (3) is plugged in SGD is presented in Algorithm 1.

**Algorithm 1:** Pseudo code SGIS

**Input** $\theta^{(0)} \in \mathbb{R}^d$ initial point, $T \geq 1$ number of iterations, $\eta > 0$ learning rate, $N \geq 1$ number of Monte-Carlo samples.
**Output** $\theta^{(0)}, \ldots, \theta^{(T-1)}$
**for** $t = 0 \ldots T-1$ **do**
  Sample $i(t) \sim \text{Unif}\{1, \ldots, n\}$
  Sample $\mathbf{V}_k \sim \pi^{(t)} (1 \leq k \leq N)$
  Compute $\widehat{g}^{(t)}$ as in Eq. (3)
  Update $\theta^{(t+1)} = \theta^{(t)} + \eta \widehat{g}^{(t)}$
**end**

# 3 Convergence guarantees

A differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be $L-$smooth for $L \geq 0$ if for all $\theta, \theta' \in \mathbb{R}^d, \|\nabla_\theta f(\theta) - \nabla_\theta f(\theta')\| \leq L \|\theta - \theta'\|$. The following theorem states sufficient conditions to ensure convergence of SGD.

**Theorem 3.1.** *[Ajalloeian and Stich [2021]] Let $\epsilon > 0, \theta^{(0)} \in \mathbb{R}^d$ and assume $\ell$ is $L-$smooth. If there exists $\xi, \sigma > 0$ such that for all $t \geq 1$, $G^{(t)}$ is chosen such that*

$$\mathbb{E}_{G^{(t)}} \left[ \left\| \widehat{g}^{(t)} - \nabla_\theta \ell \left( \theta^{(t)} \right) \right\|^2 \right] \leq \sigma$$
$$\left\| \mathbb{E}_{G^{(t)}} \left[ \widehat{g}^{(t)} \right] - \nabla_\theta \ell \left( \theta^{(t)} \right) \right\|^2 \leq \xi,$$

*then the sequence $\left(\theta^{(t)}\right)_{0 \leq t \leq T-1}$ defined by Algorithm 1 with $\eta = \min\left(\frac{1}{L}, \frac{\epsilon+\zeta}{2L\sigma}\right)$ and $T \geq K\left(\frac{1}{\epsilon+\zeta} + \frac{\sigma}{\epsilon^2+\zeta^2}\right)$ for some $K > 0$ satisfies*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \nabla_\theta \ell \left( \theta^{(t)} \right) \right\|^2 \right] \leq \tilde{K} \left( \epsilon + \zeta \right), \tag{4}$$

*for some $\tilde{K} > 0$.*

The bias of the suggested estimator 3 can be monitored thanks to the following theorem.

**Theorem 3.2.** *[Agapiou et al. [2017]] For all $t \geq 0$ and $1 \leq i \leq n$ we have*

$$\mathbb{E}_{G^{(t)}} \left[ \left\| \widehat{g}^{(t)} - \nabla_\theta \ell_{i(t)} \left( \theta^{(t)} \right) \right\|^2 \right] \leq \frac{M_{\pi^{(t)}, \theta^{(t)}}}{N} \tag{5}$$

*and*

$$\left\|\mathbb{E}_{G^{(t)}}\left[\widehat{g}^{(t)}\right] - \nabla_\theta \ell_{i(t)}\left(\theta^{(t)}\right)\right\|^2 \leq \frac{\tilde{M}_{\pi^{(t)},\theta^{(t)}}}{N} \tag{6}$$

*with* $M_{\pi^{(t)},\theta^{(t)}}$ *and* $\tilde{M}_{\pi^{(t)},\theta^{(t)}}$ *are constant detailed in the appendix.*

In the following, we assume that $\theta \in \mathcal{X}$ with $\mathcal{X}$ a compact convex subset of $\mathbb{R}^d$. Consider $\pi$ a distribution on $\mathbb{R}^q$ and $\psi : \mathbb{R}^q \mapsto \mathbb{R}^p$, we denote $\psi \in \mathcal{L}^r(\pi)$ if

$$\mathbb{E}_\pi\left[\|\psi(W)\|_1^r\right] < \infty.$$

**Corollary 3.3.** *[Convergence guarantees]*

*Let* $1 \leq i \leq n$ *a parametric family of positive density on* $\mathbb{R}^q$ *and* $\lambda_i \in \mathbb{R}_+$ *such that* ${}^{p_{\theta^{(t)}}(\cdot|\mathbf{Y}_i)}/_{\pi(\cdot,\theta^{(t)},i)} \leq \lambda_i$ *for all* $t \geq 0$. *If* $\mathbb{E}_{\pi(\cdot,\theta,i)}\left[\|\nabla_\theta \log\left(p_\theta(\mathbf{Y}_i, \mathbf{W})\right)\|_1^4\right]$ *is finite then assumptions of Theorem 3.1 are verified for Algorithm 1 if the iterates* $(\theta^{(t)})_{t\geq 1}$ *are assumed to belong to the compact convex subset* $\mathcal{X}$. *Moreover, the bias asymptotically vanishes as* $\zeta = \frac{\alpha}{N}$ *for some* $\alpha > 0$.

# 4 Discussion

Theorem 3.1 can be applied only if Theorem 3.2 is valid, which requires the iterates $(\theta^{(t)})_{t\geq 1}$ remains in the compact $X$, but cannot be proved without loss of generality. An alternative would be to consider a projection step:

$$\theta^{(t+1)} = \theta^{(t)} - P_{\mathcal{X}}\left(\theta^{(t)} - \eta\widehat{g}^{(t)}\right)$$

where $P_{\mathcal{X}}(\cdot)$ denotes the projection on $\mathcal{X}$, but this does not fall into the setting of Theorem 3.1 so that an adjustment of this Theorem would be preferable and is currently under investigation.

Simulations have been performed and comparison with Chiquet et al. [2018] are promising.

# References


Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of sgd with biased gradients, 2021.

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost, 2017.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674 – 2698, 2018.

# Session groupe MALIA

# High probability and risk-averse guarantees for a stochastic accelerated primal-dual method[1]

Yassine Laguel [1]

[1] *Université Côte d'Azur, France, yassine.laguel@univ-cotedazur.fr*

**Résumé.** Les problèmes point-selle apparaissent dans de nombreuses applications, allant de l'apprentissage robuste à la théorie des jeux, en passant par les problèmes d'équité statistique. Dans ce projet, nous étudions les propriétés de convergence de `SAPD`, un algorithme du premier ordre pour les problèmes point-selle stochastiques fortemement monotones. Nous démontrons la convergence à une vitesse accéléré de l'algorithme, en haute probabilités et pour différentes mesures de risque convexes. Pour les problèmes quadratiques sous perturbations gaussiennes, nous dérivons des formules analytiques sur la matrice de covariance limite des iterées ainsi que des bornes inférieures de complexité qui montrent que notre analyse générale est optimale. Nous illustrons nos résultats avec des expériences numériques sur des jeux à somme nulle et des problèmes d'apprentissage robustes.

**Mots-clés.** Problèmes point-selles, optimisation stochastique, preuve en grande probabilité, mesures de risque.

**Abstract.** Stochastic saddle point problems arise in many applications, ranging from distributionally robust learning to game theory and fairness in machine learning. We investigate the stochastic accelerated primal-dual algorithm for strongly-convex-strongly-concave (SCSC) saddle point problems. Our algorithm offers optimal complexity in several settings and we provide high probability guarantees for convergence to a neighbourhood of the saddle point. For quadratic problems under Gaussian perturbations, we derive analytical formulas for the limit covariance matrix together with lower bounds that show that our general analysis for SCSC problems is tight. Our risk-averse convergence analysis characterises the trade-offs between bias and risk in approximate solutions. We present numerical experiments on zero-sum games and robust learning problems..

**Keywords.** Saddle point problems, Stochastic optimization, High probability analyses, Risk measures.

**Link to paper** : `https://arxiv.org/pdf/2304.00444.pdf`

# 1 SAPD: A robust accelerated algorithm for stochastic min-max problems

## 1.1 Problem setting

We consider strongly convex/strongly concave (SCSC) saddle point problems of the form:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\mathcal{L}(x,y)\triangleq f(x)+\Phi(x,y)-g(y), \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional Euclidean spaces, $f:\mathcal{X}\to\mathbb{R}\cup\{+\infty\}$ (resp. $g:\mathcal{Y}\times\mathbb{R}\cup\{+\infty\}$) is closed and $\mu_x$-strongly convex (resp. $\mu_y$-strongly convex), and $\Phi:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is a smooth convex-concave function. Specifically, we assume that the partial gradients of $\nabla_x\Phi$ and $\nabla_y\Phi$ of $\Phi$ satisfy

$$\begin{aligned}\|\nabla_x\Phi(x,y)-\nabla_x\Phi(\bar{x},\bar{y})\|&\leq L_{x,x}\|x-\bar{x}\|+L_{x,y}\|y-\bar{y}\|,\\ \|\nabla_y\Phi(x,y)-\nabla_y\Phi(\bar{x},\bar{y})\|&\leq L_{y,x}\|x-\bar{x}\|+L_{y,y}\|y-\bar{y}\|,\end{aligned}$$

for all $(x,y),(\bar{x},\bar{y})\in\mathbf{dom}\,f\times\mathbf{dom}\,g$. We further assume that these gradient are available only through light-tail stochastic estimates $\widetilde{\nabla}_x\Phi$ and $\widetilde{\nabla}_y\Phi$: the gradient noise terms $\Delta_k^x=\widetilde{\nabla}_x\Phi(x_k,y_{k+1})-\nabla_x\Phi(x_k,y_{k+1})$ and $\Delta_k^y=\widetilde{\nabla}_y\Phi(x_k,y_k)-\nabla\Phi(x_k,y_k)$ are assumed norm-subGaussian [1] with respective proxies $\delta_x,\delta_y>0$. Such setting arises frequently in large-scale optimization and machine learning applications where the gradients are estimated from either streaming data or from random samples of data (see e.g. [2, 3]).

## 1.2 The SAPD Algorithm

We consider the Stochastic Accelerated Primal Dual algorithm (SAPD), introduced in [4], which interleaves stochastic proximal gradient ascent steps with respect to $y$ and stochastic proximal gradient descent steps with respect to $x$. Precisely, for $k\geq 0$, it maintains the iterates, $x_k, y_k$, and $s_k$ as

$$\begin{aligned}\tilde{\mathrm{s}}_k&\leftarrow\tilde{\nabla}_{\mathrm{y}}\Phi\left(x_k,y_k,\omega_k^y\right)+\theta\left(\tilde{\nabla}_{\mathrm{y}}\Phi\left(x_{k+1},y_{k+1},\omega_{k+1}^y\right)-\tilde{\nabla}_{\mathrm{y}}\Phi\left(x_k,y_k\right)\right)\\ y_{k+1}&\leftarrow\mathrm{prox}_{\sigma g}\left(y_k+\sigma\tilde{\mathrm{s}}_k\right)\\ x_{k+1}&\leftarrow\mathrm{prox}_{\tau f}\left(x_k-\tau\tilde{\nabla}_{\mathrm{x}}\Phi\left(x_k,y_{k+1}\right)\right).\end{aligned} \tag{2}$$

where $\tau$ and $\sigma$ denote respectively stepsize parameters with respect to $x$ and $y$, and $\theta\in(0,1)$ a momentum parameter. The use of momentum on the dual variable, as displayed in the previous equation, allows for provably accelerated convergence of SAPD toward an approximate solution of (1).

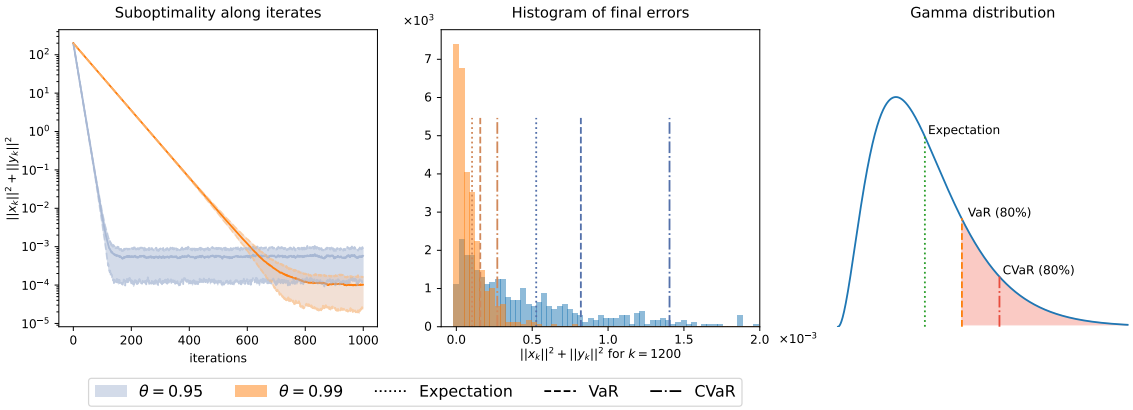



**Figure 1:** **(Left)** Convergence of SAPD on the saddle point problem $\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} x^2/2 + xy + y^2/2$, initialized at $x_0 = y_0 = 10$ with momentum parameters $\theta = 0.95$ and $\theta = 0.99$. **(Middle)** Histogram of the distribution of the SAPD iterates $(x_n, y_n)$ after $n = 1000$ iterations for 500 runs, with corresponding momentum parameters $\theta = 0.95$ and $\theta = 0.99$. **(Right)** Illustration of the expectation $\mathbb{E}(X)$, $p$-th quantile $(\mathrm{VaR}_{1-p}(X))$ and $\mathrm{CVaR}_p(X)$ for $p = 80\%$ and $X$ a gamma-distributed random variable with shape parameter 3 and scale parameter 5.

# 2 Main contributions

## 2.1 High Probability Bounds

Our main contribution lies in providing the first analysis of an accelerated algorithm for SCSC problems with high probability guarantees, where our bounds reflect the accelerated decay of the initialization bias scaling linearly with the condition number of the problem. More specifically, our high-probability bounds imply that given target accuracy $\varepsilon > 0$, `SAPD`, with a proper choice of parameters that we explicit, can generate a solution $(x_n, y_n)$ that satisfies $\mu_x\|x_n - x^\star\|^2 + \mu_y\|y_n - y^\star\|^2 \leq \varepsilon$ with probability $p \in (0,1)$ after

$$n = \mathcal{O}\left(\left[\frac{L_{x,x}}{\mu_x} + \frac{L_{y,x}}{\sqrt{\mu_x\mu_y}} + \frac{L_{y,y}}{\mu_y} + \left(1 + \frac{L_{x,y}}{L_{y,x}} + \frac{L_{x,y}^2}{L_{y,x}^2}\right)\max\left(\frac{\delta_x^2}{\mu_x}, \frac{\delta_y^2}{\mu_y}\right)\frac{1 + \log\left(\frac{1}{1-p}\right)}{\varepsilon}\right]\log\left(\frac{(1 + \frac{L_{x,y}^2}{L_{y,x}^2})\mathcal{W}_0}{\varepsilon}\right)\right) \tag{3}$$

iterations where $\mathcal{W}_0 \triangleq \mu_x \|x_0 - x^\star\|_2^2 + \mu_y \|y_0 - y^\star\|_2^2$.

## 2.2 Refined analysis for quadratics

We provide an in-depth analysis of the behavior of `SAPD` on a class of quadratic problems subject to i.i.d. isotropic Gaussian noise where we can characterize the behavior of the distribution of the iterates explicitly. In particular, we derive an analytical formula for the limiting covariance matrix of `SAPD`'s iterates. This is achieved through the solving of an intricate Lyapunov equation parameterized by the stepsize and momentum parameters of the algorithm. We leverage this formula to demonstrate the tightness of our high probability bounds for general SCSC problems.

## 2.3 Further risk-averse bounds

We finally provide finite-time risk guarantees for the convergence of `SAPD`, where we measure the risk in terms of the Conditional Value at Risk [5], the Entropic Value at Risk [6] and $\chi^2$-divergence based risk measure of the distance to the saddle point. To our knowledge, these are the first risk-averse guarantees that quantify the risk associated with an *approximate* solution generated by a primal-dual algorithm for saddle point problems.

# References


[1] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subGaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

[2] Mert Gürbüzbalaban, Andrzej Ruszczyński, and Landi Zhu. A stochastic subgradient method for distributionally robust non-convex and non-smooth learning. *Journal of Optimization Theory and Applications*, 194(3):1014–1041, 2022.

[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[4] Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points, 2023. Available at https://arxiv.org/pdf/2111.12743.

[5] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[6] Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155:1105–1123, 2012.

# LIER LA THÉORIE PAC-BAYÉSIENNE AUX MINIMA PLATS

Maxime Haddouche[1] & Paul Viallard [2],
& Umut Simsekli[3], & Benjamlin Guedj[4]

[1] *Inria London & Université de Lille, France, maxime.haddouche@inria.fr*
[2] *Inria & Université de Rennes, France, paul.viallard@inria.fr*
[3] *Inria Paris, France, umut.simsekli@inria.fr*
[4] *Inria London & University College London, UK, benjamin.guedj@inria.fr*

**L'apprentissage automatique moderne implique généralement des prédicteurs dans un contexte sur-paramétré (nombre de paramètres entraînés supérieur à la taille de l'ensemble de données), et leur apprentissage donne non seulement de bonnes performances sur les données d'entraînement, mais également une bonne capacité de généralisation. Ce phénomène remet en question de nombreux résultats théoriques et reste un problème ouvert. Pour parvenir à une meilleure compréhension, nous présentons dans cet exposé de nouvelles limites de généralisation impliquant des termes de gradient. Pour ce faire, nous combinons la théorie PAC-Bayésienne avec les inégalités de Poincaré et Log-Sobolev, évitant ainsi une dépendance explicite à la dimension de l'espace des prédicteurs. Ces résultats mettent en évidence l'influence positive des *minima plats* (étant des minima avec un voisinage minimisant presque le problème d'apprentissage) sur les performances de généralisation, impliquant directement les bénéfices de la phase d'optimisation.**

**PAC-Bayes, Statistical Learning, Flat Minima, Poincaré and Log-Sobolev Inequalities ...**

**Abstract.**Modern machine learning usually involves predictors in the overparametrised setting (number of trained parameters greater than dataset size), and their training yield not only good performances on training data, but also good generalisation capacity. This phenomenon challenges many theoretical results, and remains an open problem. To reach a better understanding, we present in this talk novel generalisation bounds involving gradient terms. To do so, we combine the PAC-Bayes toolbox with Poincaré and Log-Sobolev inequalities, avoiding an explicit dependency on dimension of the predictor space. Those results highlight the positive influence of *flat minima* (being minima with a neighbourhood nearly minimising the learning problem as well) on generalisation performances, involving directly the benefits of the optimisation phase.

# 1 Introduction

Understanding generalisation in modern machine learning problems has been a major challenge in learning theory. The goal here is to upper-bound the so-called *generalisation error* that is gap between the population and empirical risks, $\mathrm{R}_{\mathcal{D}}(h) - \hat{\mathrm{R}}_{\mathcal{S}_m}(h)$, where $h \in \mathbb{R}^d$ is the parameters of a predictor, $\mathrm{R}_{\mathcal{D}} := \mathbb{E}_{\mathbf{z}\sim\mu}[\ell(h, \mathbf{z})]$ is the population risk, $\mathcal{D}$ is an unknown data distribution, $\ell$ is a

loss function, $\hat{\mathrm{R}}_{\mathcal{S}_m} := \frac{1}{m}\sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$, and finally $\mathcal{S}_m := \{\mathbf{z}_1, \cdots, \mathbf{z}_m\}$ is a dataset with each $\mathbf{z}_i$ is independent and identically distributed (*i.i.d.*) with $\mathcal{D}$. Dating back to Hochreiter and Schmidhuber [1997], it has been hypothesised that the notion of 'flatness' (or sometimes equivalently referred to as 'sharpness') has tight links with the generalisation error: among the minima (belonging to $\hat{\mathrm{R}}_{\mathcal{S}_m}$) that is found by the learning algorithm, the 'flatter' the minimum is, the lower is the generalisation error. While the initial flatness notion was (vaguely) defined through low Kolmogorov complexity, there is no single formal definition of 'flatness'. Hence, several flatness notions have been considered, which typically are based on the second-order derivatives of the empirical risk around the local minimum found by the algorithm, such as $\mathrm{trace}(\nabla^2 \hat{\mathrm{R}}_{\mathcal{S}_m}(h))$, see *e.g.*, Jastrzebski et al. [2017], Wen et al. [2023].

While there have been several attempts to link some form of flatness to generalisation in a mathematically rigorous way [Neyshabur et al., 2017, Petzka et al., 2021, Yue et al., 2023, Andriushchenko et al., 2023], mainly in the framework of 'sharpness aware minimisation' [Foret et al., 2020], it has been recently shown that flat minima do not always imply good generalisation. In fact, there exist scenarios such that the flattest minima achieve the worst generalisation performance compared to non-flat ones [Wen et al., 2023]. In this study, we aim at developing novel links between flatness and the generalisation error from a PAC-Bayesian perspective [see *e.g.*, Guedj, 2019, Hellström et al., 2023, Alquier, 2024]. Denoting by $\mathrm{Q}$, the probability distribution of the algorithm output $h$ (or the output of a learning algorithm), we identify sufficient conditions on $\mathrm{Q}$ such that flatness always implies good generalisation. More precisely, we make the following contributions:

- We show that, when $\mathrm{Q}$ satisfies the Poincaré inequality and a technical condition that we identify, we can obtain a 'fast-rate' generalisation bound that diminishes with rate $1/m$ (rather than $1/\sqrt{m}$) and mainly contains two terms:
  - (i) The flatness term: $\mathbb{E}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m} \|\nabla_h \ell(h, \mathbf{z}_i)\|^2\right]$. This term is directly linked to the Hessian of the loss $\ell$, due to the connection between the Fisher information and the Hessian of the loss Bickel and Doksum [2015]. For instance, under certain conditions, it can be shown that $\mathrm{trace}(\nabla^2 \hat{\mathrm{R}}_{\mathcal{S}_m}(h)) = \frac{2}{m}\sum_{i=1}^{m} \|\nabla_h \ell(h, \mathbf{z}_i)\|^2$ [Wen et al., 2023, Lemma 4.1].
  - (ii) The classical PAC-Bayesian complexity term $KL(\mathrm{Q}, \mathrm{P})$, where $KL$ denotes the Kullback-Leibler divergence and $\mathrm{P}$ is data-independent 'prior' distribution.
- We then further analyse the term $KL(\mathrm{Q}, \mathrm{P})$. We show that, when $\mathrm{Q}$ is a Gibbs distribution, *i.e.*, $\mathrm{Q}(h) \propto \exp(-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}(h))\mathrm{P}(h)$ for some $\gamma > 0$ and $\mathrm{P}$ satisfies a log-Sobolev inequality, the generalisation error can be controlled *solely* by the term: $\gamma^2 c_{LS}(\mathrm{P})\, \mathbb{E}_{h\sim\mathrm{Q}}[\|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2]$, where $c_{LS}(\mathrm{P})$ denotes the log-Sobolev constant of the prior $\mathrm{P}$.

Our results shed further light on the impact of the flatness of the minima over the generalisation error: when the learning algorithm ensures a sufficiently regular distribution over the parameters, the generalisation error can be directly controlled by the flatness of the region found by the algorithm.

## 2 Preliminaries

**Framework.** We consider a predictor set $\mathcal{H} \subseteq \mathbb{R}^d$ equipped with a norm $\|.\|$, a data space $\mathcal{Z}$ and the space of distributions over $\mathcal{H} \in \mathcal{M}(\mathcal{H})$. We also consider a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. We assume that we have access to a *i.i.d.* dataset $\mathcal{S} = (\mathbf{z}_i)_{i \geq 1} \in \mathcal{Z}^{\mathbb{N}}$ with associated distribution $\mathcal{D}$. For each $m \geq 1$, we define $\mathcal{S}_m := \{\mathbf{z}_1, \cdots, \mathbf{z}_m\}$. In PAC-Bayes learning, we construct a data-driven posterior distribution $Q \in \mathcal{M}(\mathcal{H})$ with respect to a prior distribution $P$. To assess the generalisation ability of a predictor $h \in \mathcal{H}$, we define the *population risk* to be $\mathrm{R}_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{z} \sim \mu}[\ell(h, \mathbf{z})]$ and for each $m$, its empirical counterpart $\hat{\mathrm{R}}_{\mathcal{S}_m}(h) := \frac{1}{m} \sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$. As PAC-Bayes focuses on elements of $\mathcal{M}(\mathcal{H})$, we also define the expected risk and empirical risks for $Q \in \mathcal{M}(\mathcal{H})$ as $\mathrm{R}_{\mathcal{D}}(Q) := \mathbb{E}_{h \sim Q}[\mathrm{R}_{\mathcal{D}}(h)]$ and $\hat{\mathrm{R}}_{\mathcal{S}_m}(Q) := \mathbb{E}_{h \sim Q}[\hat{\mathrm{R}}_{\mathcal{S}_m}(h)]$. PAC-Bayes bounds usually aim at controlling the *expected generalisation error (or gap)* for each dataset size $m$, i.e., $\Delta_{\mathcal{S}_m}(Q) := \mathrm{R}_{\mathcal{D}}(Q) - \hat{\mathrm{R}}_{\mathcal{S}_m}(Q)$.

**Background on Poincaré and log-Sobolev inequalities.** In this work, we exploit Poincaré and log-Sobolev inequalities in the PAC-Bayes framework. We first recall the definition of Poincaré and log-Sobolev inequalities. To do so, for a fixed distribution $\mathrm{Q}$, we define the *Sobolev space of order* 1 on $\mathbb{R}^d$ as follows:

$$\mathrm{H}^1(\mathrm{Q}) := \left\{ f \in \mathrm{L}^2(\mathrm{Q}) \cap \mathrm{D}_1(\mathbb{R}^d) \mid \|\nabla f\| \in \mathrm{L}^2(\mathrm{Q}) \right\},$$

where $\mathrm{D}_1(\mathbb{R}^d)$ denotes the set of derivable functions $f : \mathbb{R}^d \to \mathbb{R}$.

**Definition 1 (Poincaré and Logarithmic Sobolev inequalites)** *A measure* $\mathrm{Q}$ *satisfies a* Poincaré inequality *with constant* $c_P(\mathrm{Q})$ *if for all function* $f \in \mathrm{H}^1(\mathrm{Q})$ *we have*

$$\mathrm{Var}_{\mathrm{Q}}(f) \leq c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{h \sim \mathrm{Q}} \left[ \|\nabla f(h)\|^2 \right],$$

*where* $\mathrm{Var}_{\mathrm{Q}}(f) = \mathbb{E}_{h \sim \mathrm{Q}} \left[ f(h) - \mathbb{E}_{h \sim \mathrm{Q}}[f(h)] \right]^2$ *is the* variance *of* $f$ *w.r.t.* $\mathrm{Q}$*. We then say that* $\mathrm{Q}$ *is Poincaré with constant* $c_P(\mathrm{Q})$*, or that* $\mathrm{Q}$ *is* $\texttt{Poinc}(c_P)$*. Also,* $\mathrm{Q}$ *satisfies a* log-Sobolev inequality *with constant* $c_{LS}(\mathrm{Q})$ *if for all function* $f \in \mathrm{H}^1(\mathrm{Q})$ *we have*

$$\mathop{\mathbb{E}}_{h \sim \mathrm{Q}} \left[ f^2(h) \log \left( \frac{f^2(h)}{\mathbb{E}_{h \sim \mathrm{Q}} \left[ f^2(h) \right]} \right) \right] \leq c_{LS}(\mathrm{Q}) \mathop{\mathbb{E}}_{h \sim \mathrm{Q}} \left[ \|\nabla f(h)\|^2 \right],$$

*where the term on the left hand side is the* entropy *of* $f^2$*, denoted as* $\mathrm{Ent}_Q(f^2)$*. We then say that* $\mathrm{Q}$ *is log-Sobolev with constant* $c_{LS}(\mathrm{Q})$*, or that* $\mathrm{Q}$ *is* $\texttt{L-Sob}(c_{LS})$*.*

The class of Gaussian distributions is an important particular case of distributions satisfying both Poincaré and log-Sobolev inequalities A gaussian $\mathrm{Q}$ with covariance matrix $\Sigma_{op}$, $\mathrm{Q}$ is $\texttt{L-Sob}(c_{LS})$ with constant $c_{LS}(\mathrm{Q}) = 2\|\Sigma\|_{op}$ and also $\texttt{Poinc}(c_{LS})$ with constant $c_{LS}(\mathrm{Q}) = \|\Sigma\|_{op}$, where $\|.\|_{op}$ denotes the operator norm. Also, we focus on specific posterior distributions called *Gibbs posteriors, or Gibbs distributions*. For a fixed loss $\ell$ and dataset $\mathcal{S}_m$, the Gibbs posterior, *w.r.t.* prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, risk $\hat{\mathrm{R}}_{\mathcal{S}_m}$ and *inverse temperature* $\gamma > 0$.is defined as $\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}$ such that $d\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}(h) \propto \exp(-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}(h)) dP(h)$. Gibbs posteriors are a class of closed-form solutions for relaxation of Catoni [2007, Theorem 1.2.6] stated, for instance, in Alquier et al. [2016, Theorem 4.1]. Theorem 2 shows that when the prior and the loss satisfies a few properties, then the associated Gibbs posterior is $\texttt{L-Sob}(c_{LS})$.

**Proposition 2** *Assume that* $\mathrm{P}$ *is a probability measure on* $\mathbb{R}^d$ *such that* $dP(h) \propto \exp(-V(x))$ *with* $V$ *a smooth function such that* $Hess(V) \succeq \frac{2}{c_{LS}(\mathrm{P})}\mathrm{Id}$. *Assume that* $\ell = \ell_1 + \ell_2$ *with* $\ell_1$ *convex, twice differentiable and* $\ell_2$ *bounded. Then for any* $\gamma > 0$*, the Gibbs posterior* $\mathrm{Q} = \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}$ *is* $\texttt{L-Sob}(c_{LS})$ *with constant* $c_{LS}(\mathrm{Q}) = c_{LS}(\mathrm{P})\exp\left(4\|\ell_2\|_\infty\right)$.

Theorem 2 applies, *e.g.*, when $\mathrm{P}$ is a Gaussian prior $\mathrm{P} = \mathcal{N}(\mu_\mathrm{P}, \Sigma_\mathrm{P})$. Notice that in this case $c_{LS}(\mathrm{P}) = 2\|\Sigma_\mathrm{P}\|_{op}$. This property is a straightforward application of Chafaï [2004, Corollary 2.1] with Guionnet and Zegarlinksi [2003, Property 2.6]. Finally, notice that satisfying a log-Sobolev inequality is stronger than satisfying a Poincaré one. This is stated for instance in Ledoux [2006, Proposition 2.1].

# 3 Reaching a flat minimum allows Poincaré posteriors to generalise well

**Fast rate PAC-Bayes bounds for heavy-tailed losses** In order to obtain fast rates, *i.e.*, bounds converging to zero faster than $1/\sqrt{m}$, we exploit the notion of flat minimum (where the loss takes a small value in the neighbourhood of the minimum). Indeed, in an overparametrised setting such as neural networks, it is likely to obtain such a minimum once the optimisation phase has been performed, as there are much more parameters than training data. We exploit this flatness property within PAC-Bayes bounds through the gradient norm $\|\nabla_h \ell(., \mathbf{z})\|$ of the loss *w.r.t.* the predictor $h$ for any $\mathbf{z}$. In this section, we consider posterior distributions $\mathrm{Q}$ being $\texttt{Poinc}(c_P)$. This assumption covers the important case of Gaussian measures as well as all measures satisfying a log-Sobolev inequality. We focus on PAC-Bayes bound holding for distributions $\mathrm{Q}$ satisfying a particular assumption involving the data distribution $\mathcal{D}$ (contrary to many PAC-Bayes bounds holding for all $\mathrm{Q}$). We then define the *error* of $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ for any datum $\mathbf{z} \in \mathcal{Z}$ as $\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z}) := \mathbb{E}_{h\sim\mathrm{Q}}[\ell(h, \mathbf{z})]$ and identify Assumption 3 to later involve flat minima.

**Assumption 3** *We say that* $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ *is* quadratically self-bounded *with respect to* $\ell$ *and constant* $C > 0$ *(namely* $\texttt{QSB}(\ell, C)$*) if*

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})^2\right] \leq C\mathrm{R}_\mathcal{D}(\mathrm{Q})\,(= C\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})\right])$$

Assumption 3 is a relaxation of boundedness, as if $\ell \in [0, C]$ then it is $\texttt{QSB}(\ell, C)$. It is an alternative to the bounded expected variance assumption in anytime-valid PAC-Bayes bounds [Haddouche and Guedj, 2023, Chugg et al., 2023]. An issue with such boundedness assumption is that it has to hold for all posteriors, including those providing poor generalisation performances. This is avoided by the $\texttt{QSB}$ assumption which intricate the properties of $\mathcal{D}, \ell$ and $\mathrm{Q}$. Finally, we interpret $C$ as a contraction constant attenuating, on average, the local expansion (governed by variances of $\mathrm{Q}$, and $\mathcal{D}$) of the loss around the mean of $\mathrm{Q}$. Exploiting the PAC-Bayes supermartingales bounds of Haddouche and Guedj [2023], Chugg et al. [2023] alongside Poincaré inequality leads to the following.

**Theorem 4** *For any $C > 0$, any $\frac{2}{C} > \lambda > 0$, any data-free prior $\mathrm{P}$, any $\ell \geq 0$ and any $\delta \in [0,1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any $\mathrm{Q}$ being $\texttt{Poinc}(c_P)$, $\texttt{QSB}(\ell, C)$ and $\ell(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,*

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq \frac{1}{1-\frac{\lambda C}{2}} \left( \hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda m} \right) + \frac{\lambda}{2-\lambda C} c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}} \left[ \mathop{\mathbb{E}}_{h\sim\mathrm{Q}} \left( \|\nabla_h \ell(h,\mathbf{z})\|^2 \right) \right].$$

This theorem shows that, for any posterior being $\texttt{QSB}$ *w.r.t.* the distribution $\mathcal{D}$, fast rates are achievable as long as $\hat{\mathrm{R}}_{\mathcal{S}_m} \approx 0$, and expected gradients are vanishing. While the first condition is often involved for deep neural networks in the overparametrised setting, the second holds if a flat minimum has been reached through the optimisation process. Then, taking $\lambda = 1/C$ ensures an anytime-valid PAC-Bayesian bound with a fast rate of $1/m$. Otherwise, for a fixed $m$, taking $\lambda = m^{-\alpha}/C$, $\alpha \in [0; 1/2]$ allows to adapt the convergence speed *w.r.t.* the behaviour of the gradients. In the case of constant gradients, we recover a convergence rate of $1/\sqrt{m}$, matching Alquier et al. [2016, Theorem 4.1].

**On the role of flat minima in PAC-Bayes learning.** Theorem 4 suggests that, in order to attain good generalisation ability, the mean of $\mathrm{Q}$ has to be close from two minima: *(i)* on $\hat{\mathrm{R}}_{\mathcal{S}_m}$ in order to make $\hat{\mathrm{R}}_{\mathcal{S}_m}$ small, and *(ii)* on $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\|\nabla_h \ell(h,\mathbf{z})\|^2]$ to make the gradients small. The variance of $\mathrm{Q}$ has to fit the flatness of those minima, the flatter they are, the larger the variance in order to shrink the expected terms on the right-hand-side of Theorem 4. Finally, the KL term invites, *e.g.* for Gaussian distributions, to consider high variances, hence flat minima to maintain a small value of the bound.

**A focus on $C$.** Taking $\lambda = 1/C$ in Theorem 4 attenuates the impact of the prior distribution and amplifies the gradient term. Then, a small $C$ is desirable when working with flat minima to attenuate an ill-designed prior.

**High probability bounds with fast rates, a paradox?** Grunwald et al. [2021, page 7] showed that, for a trivial $\mathcal{H} = \{h\} \subset \mathbb{R}^d$, for any loss, any *i.i.d.* dataset $\mathcal{S}_m$ with variance $\sigma^2$, we have asymptotically, with probability at least $\alpha$, for a constant $C_\alpha$ depending on $\alpha$ and $\mathcal{N}(\mathbf{0}, \mathrm{Id})$, we have $\mathrm{R}_{\mathcal{D}}(h) \geq \hat{\mathrm{R}}_{\mathcal{S}_m}(h) + C_\alpha \frac{\sigma^2}{\sqrt{m}}$. Is it paradoxical with Theorem 4? The answer is no: the bound in Grunwald et al. [2021] gives an asymptotic lower bound on the convergence of $\hat{\mathrm{R}}_{\mathcal{S}_m}(h)$ to $\mathrm{R}_{\mathcal{D}}(h)$. Theorem 4 informs us on how $\mathrm{R}_{\mathcal{D}}$ is getting closer from $\frac{1}{1-\lambda/2}\hat{\mathrm{R}}_{\mathcal{S}_m}$ which converges to $\frac{1}{1-\lambda/2}\mathrm{R}_{\mathcal{D}} > \mathrm{R}_{\mathcal{D}}$ as the loss is non-negative. Theorem 4 then show the existence of a 'transition regime' involving a fast rate. Once $\frac{1}{1-\lambda/2}\hat{\mathrm{R}}_{\mathcal{S}_m}$ is reached, the clower bound of Grunwald et al. [2021] ensures an asymptotic regime with slow convergence rate. Note that such transition regimes already appeared in the literature in Tolstikhin and Seldin [2013], Mhammedi et al. [2019] at the cost of additional variance terms compared to Theorem 4. However, such fast rates have never been linked before to flat minima (and optimisation in general), highlighting the potential of our bound to explain the ability of deep neural networks to generalise well in the overparametrised setting ($m$ far smaller than the dimension of $\mathcal{H}$), where flat minima are likely to be reached, as studied, *e.g.*, in Dziugaite et al. [2020], showing correlations between flat minima and generalisation for various learning problems.

It is possible to go beyond the QSB assumption. This comes at the cost of an upper bound on $\mathrm{R}_{\mathcal{D}}$ as well as a supplementary Poincaré assumption on $\mathcal{D}$.

**Corollary 5** *For any $C > 0$, any $\delta \in (0,1)$ any $\frac{2}{C} > \lambda > 0$, any data-free prior $\mathrm{P}$, any $\ell \geq 0$ such that, for any $\mathbf{z} \in \mathcal{Z}$, we have $\ell(.,\mathbf{z}) \in \mathrm{H}^1$ and for any $h$, the loss function $\ell(h,.)$ is $\mathcal{C}^1$ almost everywhere on $\mathcal{Z}$. If the data distribution $\mathcal{D}$ is $\mathtt{Poinc}(c_P)$, then with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any posterior $\mathrm{Q}$ being $\mathtt{Poinc}(c_P)$ with $\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$:*

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq \frac{1}{1-\frac{\lambda C}{2}}\left(\hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda m}\right) + \frac{\lambda}{2-\lambda C}\left(c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] + c_P(\mathcal{D}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left(\left\|\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\nabla_z \ell(h,\mathbf{z})]\right\|^2\right)\right).$$

Corollary 5 states that, if $\mathrm{Q}$ reached a flat minimum (meaning $\|\nabla_h \ell\|$ is small), and this minimum is robust to the training dataset (meaning $\|\nabla_{\mathbf{z}} \ell\|$ is small), then a fast rate is attainable while only requiring an upper bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{Q})$. This conclusion holds when $\mathcal{D}$ Poinc, encompassing the case of Gaussian mixtures [Schlichting, 2019], which can approximate any smooth density [as recalled in Gat et al., 2022]. However, the Poincaré constant of a general mixture is not known, and the upper bound of Schlichting [2019] scales with the number of components, involving potentially high $\chi^2$ divergences.

**Towards fully empirical bound for gradient-Lipschitz functions.** In this section, we assume the loss $\ell$ is such that, for any $\mathbf{z} \in \mathcal{Z}$, the gradient $\nabla_h \ell(.,\mathbf{z})$ is $G$-Lipschitz, which is often considered for convergence bounds in optimisation. A large part of high-probability PAC-Bayes bounds are fully empirical: this has numerous advantages including in-training numerical evaluation of generalisation as well as novel PAC-Bayesian algorithms, minimising such empirical bounds; see [Dziugaite and Roy, 2017, Perez-Ortiz et al., 2021, Viallard et al., 2023] among others. However, Theorem 4 and Corollary 5 are not fully empirical and thus, do not have such desirable properties. We circumvent this issue in Theorem 6.

**Theorem 6** *For any $C_1, C_2, c > 0$, any data-free prior $\mathrm{P}$, any $\ell \geq 0$ being $\mathcal{C}^2$ and any $\delta \in [0,1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any $\mathrm{Q}$ being $\mathtt{Poinc}(c_P)$ with constant $c$, $\mathtt{QSB}(\ell, C_1)$, $\mathtt{QSB}(\|\nabla_h \ell\|^2, C_2)$ and $\ell(.,\mathbf{z}), \|\nabla_h \ell\|^2(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,*

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{2c}{C_1} \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m} \|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right] + 2\left(C_1 + c\frac{4cG^2 + C_2}{C_1}\right)\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{m}.$$

Here, we showed that to attain fast rates, the QSB assumption has to be reached for both the loss and its gradient. This suggests several things on the flat minimum that has to be reached by $\mathrm{Q}$

(designed from $\hat{\mathrm{R}}_{\mathcal{S}}$): first, it needs to be close from a flat minimum of $\mathrm{R}_{\mathcal{D}}$ to satisfy the QSB assumption. Second, this minimum also ensures the contraction of the gradients. We then are able to derive an empirical generalisation bound, involving both empirical loss and gradients. Not only Theorem 6 yields, to our knowledge, the first PAC-Bayesian algorithm involving gradient terms, but also can be translated to a generalisation metric in order to understand generalisation.

# 4 Generalisation ability of Gibbs distributions with a log-Sobolev prior

One limitation of the results given in Section 3 is that the KL divergence term remains uncontrolled in general as its formulation depends on the nature of $\mathrm{P}$ and $\mathrm{Q}$. A close form exists for Gaussian distributions for instance, but this class of distribution is limiting. Perpetrating the spirit of Catoni [2007], we go beyond the Gaussian distributions to focus on the Gibbs posteriors which have naturally appeared in PAC-Bayes through the use of tools from statistical physics. We show that log-Sobolev inequalities allow us to control the KL divergence of such distributions *w.r.t.* their priors.

**Controlling the KL divergence when $\mathrm{Q}$ is a Gibbs posterior.** Theorem 7 exploits the fact that the KL divergence can be formulated as an entropy *w.r.t.* the prior distribution $\mathrm{P}$. It then shows that the KL divergence of the Gibbs posterior $\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}$ *w.r.t.* $\mathrm{P}$ is upper bounded by gradient terms as long as $\mathrm{P}$ satisfies a log-Sobolev inequality.

**Lemma 7** *For any $m$, $\mathrm{P}$ being $L\text{-}Sob(c_{LS})$, any $\ell \geq 0$ such that for any $\mathbf{z}$, $\ell(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{P})$, we have, for any $\gamma > 0$:*

$$\mathrm{KL}\left(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}, \mathrm{P}\right) \leq \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}} \left[\|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2\right].$$

The crucial message of this lemma is that, a flat minimum of $\hat{\mathrm{R}}_{\mathcal{S}}$ allows controlling the KL divergence. This message is new and independent of Section 3 which focus on flat minima reached for $\mathrm{R}_{\mathcal{D}}$. Note that in this case, the KL divergence has an explicit formulation. However it involves to calculate the exponential moment $\mathbb{E}_{h\sim\mathrm{P}}[\exp(-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m})]$ which is costly in practice. On the contrary, we only need to estimate a second-order moment over $\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}$.

**Generalisation ability of Gibbs posteriors.** When Gibbs posteriors are involved, KL divergence is controllable by a gradient term. An ideal way to conclude would be, as in Section 3 to involve Poincaré inequality. However, Gibbs posterior are not necessarily satisfying a Poincaré inequality as in Section 3, we then need to make supplementary assumptions on the loss.

**Theorem 8** *For any $C > 0$, any $\gamma > 0$, any prior $\mathrm{P}$ being $L\text{-}Sob(c_{LS})$, any $\ell \geq 0$ and any $\delta \in [0,1]$, we have the following inequalities. If $\ell \in [0,1]$, then with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, and any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:*

$$\mathrm{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}) \leq 2\left(\hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}) + \frac{\gamma^2 c_{LS}(\mathrm{P})}{4m} \mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}} \left[\|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2\right] + \frac{\log(1/\delta)}{m}\right).$$

*If $\ell = \ell_1 + \ell_2$ with $\ell_1$ convex, twice differentiable and $\ell_2$ bounded, assume that $\mathrm{P}$ satisfies the conditions of Theorem 2. Then for any $\frac{2}{C} > \lambda > 0$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, such that $\mathrm{Q}$ is $\mathtt{QSB}(\ell, C)$ and $\ell(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}})$:*

$$\mathrm{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}) \leq \frac{1}{1-\frac{\lambda C}{2}} \left( \hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}) + \frac{\gamma^2 c_{LS}(\mathrm{P})}{4\lambda m} \mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}} \left[ \|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2 \right] + \frac{\log(1/\delta)}{\lambda m} \right) + \frac{\lambda e^{4\|\ell_2\|_\infty} c_{LS}(\mathrm{P})}{4 - 2\lambda C} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}} \left[ \mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}} \left( \|\nabla_h \ell(h, \mathbf{z})\|^2 \right) \right].$$

Note that we could have derived analogous to Corollary 5 at the cost of a supplementary Poincaré assumption on $\mathcal{D}$. The influence of the inverse temperature $\gamma$ is quadratic: this is the price to pay to fit the dataset and reduce the influence of the prior. This dependency is therefore attenuated by a gradient term, small if a flat minimum on the empirical risk has been reached. This suggests that in the case of Gibbs posteriors with log-Sobolev prior, reaching a flat minima on $\hat{\mathrm{R}}_{\mathcal{S}_m}$ controls not only $\hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q})$, but also the KL divergence and this last point is not reachable when considering Poincaré distributions. The other gradient term comes from Section 3 and requires to be close from a flat minimum on $\mathrm{R}_{\mathcal{D}}$ to attain fast rates.

# 5 Conclusion

We provide novel PAC-Bayes generalisation bounds, converging faster than $1/\sqrt{m}$ when a low empirical error is reached and that expected gradients are vanishing. This conveys the message that flat minima helps generalisation. However, to complete this analysis, the crucial question is to understand how optimisation algorithms successfully reach flat minima in the overparametrised setting. This important question is left as future work.

# References


P. Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*, 2024.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 2016.

M. Andriushchenko, F. Croce, M. Müller, M. Hein, and N. Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023.

P. J. Bickel and K. A. Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. CRC Press, 2015.

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics, 2007.

D. Chafaï. Entropies, convexity, and functional inequalities, On Φ-entropies and Φ-Sobolev inequalities. *Journal of Mathematics of Kyoto University*, 44(2):325–363, 2004.

B. Chugg, H. Wang, and A. Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*, 2023. URL http://jmlr.org/papers/v24/23-0401.html.

G. K. Dziugaite and D. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

G. K. Dziugaite, A. Drouin, B. Neal, N. Rajkumar, E. Caballero, L. Wang, I. Mitliagkas, and D. M. Roy. In search of robust measures of generalization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

I. Gat, Y. Adi, A. G. Schwing, and T. Hazan. On the Importance of Gradient Norm in PAC-Bayesian Bounds. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6686e3f2e31a0db5bf90ab1cc2272b72-Abstract-Conference.html.

P. Grunwald, T. Steinke, and L. Zakynthinou. PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2217–2247. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/grunwald21a.html.

B. Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, 2019. URL https://arxiv.org/abs/1901.05353.

A. Guionnet and B. Zegarlinksi. *Lectures on Logarithmic Sobolev Inequalities*. Springer Berlin Heidelberg, 2003. doi: 10.1007/978-3-540-36107-7_1. URL https://doi.org/10.1007/978-3-540-36107-7_1.

M. Haddouche and B. Guedj. PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Transactions on Machine Learning Research*, 2023.

F. Hellström, G. Durisi, B. Guedj, and M. Raginsky. Generalization bounds: Perspectives from information theory and PAC-Bayes. *arXiv preprint arXiv:2309.04381*, 2023.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 2006.

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes Un-Expected Bernstein Inequality. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 12202–12213. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9387-pac-bayes-un-expected-bernstein-inequality.pdf.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

M. Perez-Ortiz, O. Rivasplata, E. Parrado-Hernandez, B. Guedj, and J. Shawe-Taylor. Progress in Self-Certified Neural Networks. In *NeurIPS 2021 Workshop on Bayesian Deep Learning*, 2021.

H. Petzka, M. Kamp, L. Adilova, C. Sminchisescu, and M. Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432, 2021.

A. Schlichting. Poincaré and Log-Sobolev Inequalities for Mixtures. *Entropy*, 2019. doi: 10.3390/e21010089. URL https://doi.org/10.3390/e21010089.

I. O. Tolstikhin and Y. Seldin. PAC-Bayes-Empirical-Bernstein Inequality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/a97da629b098b75c294dffdc3e463904-Paper.pdf.

P. Viallard, M. Haddouche, U. Simsekli, and B. Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *To be published in NeurIPS 2023*, 2023.

K. Wen, Z. Li, and T. Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Dkmpa6wCIx.

Y. Yue, J. Jiang, Z. Ye, N. Gao, Y. Liu, and K. Zhang. Sharpness-aware minimization revisited: Weighted sharpness as a regularization term. *arXiv preprint arXiv:2305.15817*, 2023.

# Covariance-Adaptive Least-Squares Algorithm for Stochastic Combinatorial Semi-Bandits

Julien Zhou [1] & Pierre Gaillard [2] & Thibaud Rahier [3]
& Houssam Zenati [4] & Julyan Arbel [5]

[1] *Criteo AI Lab, Inria Thoth & Statify, France, julien.zhou@inria.fr;*
[2] *Inria Thoth, France, pierre.gaillard@inria.fr ;*
[3] *Criteo AI Lab, France, t.rahier@criteo.com;*
[4] *Inria Soda & PreMeDICaL, France, houssam.zenati@inria.fr;*
[5] *Inria Statify, France, julyan.arbel@inria.fr*

**Abstract.** We address stochastic combinatorial semi-bandit problems, where a player can select from $P$ subsets of a set containing $d$ base items. This setting has been studied extensively in prior works which focused on getting acceptable regret upper bounds, despite the potentially exponentially big action space. However, most existing algorithms (e.g. `CombUCB`, `ESCB`, `OLS-UCB`) require prior knowledge on the reward distribution, like an upper bound on a sub-Gaussian proxy-variance, which is hard to estimate tightly. Their regret upper bounds also involve these hypotheses, which can make them slack depending on the instance at hand. We propose a variance-adaptive version of `OLS-UCB`, relying on an online estimation of the covariance coefficients and computing upper confidence bounds for each action available. Estimating the coefficients of a covariance matrix can be manageable in practical settings and leveraging this information with the structure of the problem results in improved regret upper bounds.

**Keywords.** Bandits; Stochastic Combinatorial Semi-Bandit; Covariance; Confidence Ellipsoid

## 1 Introduction

In sequential decision-making, the bandit framework has been studied in-depth and was instrumental to several applications. Reference books like Bubeck and Cesa-Bianchi (2012) or Lattimore and Szepesvári (2020) offer a wide perspective on the subject. In this framework, a *decision-maker* or *player* must make choices and receive associated rewards. As the player lacks prior knowledge of its an exploration-exploitation trade-off naturally arises.

We focus on the stochastic combinatorial semi-bandit framework. In this setting, the player chooses a subset of *base items* and receives a feedback for each item chosen. The corresponding action set is unfortunately potentially exponentially big and difficult to explore. However, it presents a structure that could be leveraged. The information collected by choosing different overlapping actions should be shared.

**Problem formulation.** We consider a set of $d \in \mathbb{N}^*$ *base items*, each item $i \in [d] = \{1, \dots, d\}$ yielding a stochastic reward. A *player* accesses these rewards through a set $\mathcal{A} \subseteq \{0,1\}^d$ of $P = |\mathcal{A}| \in \mathbb{N}^*$ *actions*, each corresponding to a subset of items. The player interacts with the *environment* over a sequence of $T \in \mathbb{N}^*$ *rounds*. At each round $t \in [T]$, the decision-maker chooses an action $A_t \in \mathcal{A}$, the environment samples a reward vector $Y_t \in \mathbb{R}^d$, the decision-maker observes the realization for every item contained in $A_t$, and receives their sum. The interactions between the player and the environment are summarized in Framework 1.

**Framework 1** Stochastic Combinatorial Semi-Bandit

For each $t \in \{1, \dots, T\}$:

- The player chooses an action $A_t \in \mathcal{A}$.
- The environment samples a vector of rewards $Y_t \in \mathbb{R}^d$ from a fixed unknown distribution.
- The player receives the reward $\langle A_t, Y_t \rangle = \sum_i A_{t,i} Y_{t,i}$.
- The player observes $Y_{t,i}$ for all $i \in [d]$ s.t. $A_{t,i} = 1$.

The objective of the decision-maker is to maximize the cumulative expected rewards, or equivalently to minimize the expected cumulative regret defined as:

$$\mathbb{E}[R_T] = T\langle a^*, \mu \rangle - \sum_{t=1}^{T} \mathbb{E}\big[\langle A_t, Y_t \rangle\big] = \sum_{t=1}^{T} \mathbb{E}\big[\Delta_{A_t}\big] \,, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\mathbb{R}^d$, $a^* \in \arg\max_{a \in \mathcal{A}} \langle a, \mu \rangle$ is an optimal action, and $\Delta_a = \langle a^* - a, \mu \rangle$ is the *sub-optimality gap* for action $a \in \mathcal{A}$.

**Assumptions.** We make the following assumptions on the rewards $(Y_t)_{t \in [T]}$. For all $t \in [T]$, $Y_t$ is independent of $\mathcal{F}_{t-1} = \sigma(Y_1, \dots, Y_{t-1})$ and $A_t$ is $\mathcal{F}_{t-1}-$measurable. There exists a mean reward vector $\mu \in \mathbb{R}^d$ and a positive semi-definite covariance matrix $\mathbf{\Sigma}^* \in M_d(\mathbb{R})$ such that $\mathbb{E}[Y_t] = \mu \in \mathbb{R}^d$ and $\mathrm{Var}(Y_t) = \mathbf{\Sigma}^*$. There exists a known *bounds* vector $B \in \mathbb{R}_+^d$ such that for all $i \in [d]$, $|Y_{t,i} - \mu_i| \leq B_i$.

**Contributions.** We design `OLS-UCBV`, a variance-adaptive algorithm for stochastic combinatorial semi-bandits. Compared to the existing litterature, our algorithm takes the covariance structure into account. It estimates coefficients of a covariance matrix online which can be done tightly, and efficiently. Leveraging these online estimates, `OLS-UCBV` satisfies a poly-log$(T)$ gap-dependant regret bound, with leading terms depending on the variance and the structure of the action space. We solve the limitations of Degenne and Perchet (2016) raised by Perrault et al. (2020), but with different tools yielding a computationally more efficient algorithm, and manage to get a $\sqrt{T}$ gap-free regret upper bounds contrary to the latest.

## 2 OLS-UCBV

We design OLS-UCBV, a new algorithm that efficiently leverages semi-bandit feedbacks by approximating the coefficients of the covariance matrix $\mathbf{\Sigma}^*$ online, with coefficient-wise upper bounds. OLS-UCBV satisfies the "Optimism in the face of uncertainty" (OFUL) principle of Abbasi-Yadkori et al. (2011), by deriving ellipsoidal confidence regions leveraging Bernstein-like concentration inequalities and a peeling trick.

### 2.1 Estimators for mean and covariance

**Mean estimation.** Let $a \in \mathcal{A}$, $t \in [T]$, we denote by $\mathbf{d}_a = \mathrm{diag}(a) \in M_d(\mathbb{R})$ the diagonal matrix where the non-null coefficients are the elements of the action $a$. The number of times two items $i, j \in [d]$ (with possibly $i = j$) have been chosen together after round $t$ is denoted by $n_{t,(i,j)} = \sum_{s=1}^{t} \mathbb{1}\{\{i,j\} \subseteq A_s\}$. We define $\mathbf{D}_t = \mathrm{diag}((n_{t,(i,i)})_{i\in[d]}) \in M_d(\mathbb{R})$ the diagonal matrix of item counts. Then the least-squares estimator of the mean reward vector $\mu$ using all the data from the past rounds after round $t$ is the empirical average

$$\hat{\mu}_t = \mathbf{D}_t^{-1} \sum_{s=1}^{t} \mathbf{d}_{A_s} Y_s = \mu + \mathbf{D}_t^{-1} \sum_{s=1}^{t} \mathbf{d}_{A_s} \eta_s \,, \tag{2}$$

where $\eta_s$ denotes the deviation of reward $Y_s$ from its mean $\mu$. This average yields the estimator $\langle a, \hat{\mu}_t \rangle$ for the mean reward $\langle a, \mu \rangle$, to which a well-designed optimistic bonus should be added.

**Covariance estimation.** Let $t \in \mathbb{N}^*$ and $i, j \in [d]$ such that $n_{t,(i,j)} \geq 2$. The coefficients of $\mathbf{\Sigma}^*$ can be estimated online by $\hat{\chi}_t$ with elements as follows

$$\hat{\chi}_{t,(i,j)} = \tfrac{1}{n_{t,(i,j)}} \sum_{s=1}^{t} A_{s,i} A_{s,j} \mathbb{1}\{n_{s,(i,j)} \geq 2\} (Y_{s,i} - \hat{\mu}_{s-1,i})(Y_{s,j} - \hat{\mu}_{s-1,j}) \,. \tag{3}$$

A major difference with usual batch estimators, like those used in Perrault et al. (2020) and Audibert et al., 2009, is the fact that the observed rewards are "centered" using the sample average of the past rewards at the time of the observation, instead of the whole sample average at time $t$.

### 2.2 Confidence bounds for covariance and mean

**Covariance coefficients upper confidence bound.** The following result controls the error of the online covariance estimator $\hat{\chi}_t$ presented in eq. (3).

**Proposition 1.** *Let $T \geq 3$, $\delta \in (0,1)$. Then with probability $1-\delta$, for all $t \leq T$ and $(i,j) \in [d]^2$, such that $n_{t,(i,j)} \geq 2$,*

$$|\hat{\chi}_{t,(i,j)} - \mathbf{\Sigma}^*_{i,j}| \leq \mathcal{B}_{t,(i,j)}(\delta) \,,$$

*where $\mathcal{B}_{t,(i,j)}(\delta) = 3B_iB_j \left( \frac{h_{T,\delta}}{\sqrt{n_{t,(i,j)}}} + \frac{h^2_{T,\delta}}{n_{t,(i,j)}} \log(T) \right)$ with $h_{T,\delta} = \log(5d^2T^2/\delta)$.*

This result suggests the following upper confidence bounds for $\mathbf{\Sigma}^*$ to be used into our algorithm:

$$\hat{\mathbf{\Sigma}}_{t,(i,j)} = \hat{\chi}_{t,(i,j)} + \mathcal{B}_{t,(i,j)}(\delta)\,. \tag{4}$$

It also yields $\mathcal{C}$, the high-probability event in which all the $\chi_{t,(i,j)}$ are in our confidence intervals.

$$\mathcal{C} = \left\{\forall t \in [T], (i,j) \in [d]^2 \text{ s.t. } n_{t,(i,j)} \geq 2, \qquad |\hat{\chi}_{t,(i,j)} - \mathbf{\Sigma}^*_{i,j}| \leq \mathcal{B}_{t,(i,j)}(\delta)\right\}. \tag{5}$$

**Mean upper confidence bound.** We propose an upper confidence bound for the average rewards of all actions $a \in \mathcal{A}$ at any round $t$.

Denoting by $\hat{\mathbf{\Sigma}}_t$ the covariance matrix upper bound whose coefficients are given by (4), we introduce $\hat{\mathbf{Z}}_t$ a "regularized empirical design matrix" and its "exact" counterpart $\mathbf{Z}_t$,

$$\hat{\mathbf{Z}}_t = \sum_{s=1}^{t} \mathbf{d}_{A_s} \hat{\mathbf{\Sigma}}_t \mathbf{d}_{A_s} + \mathbf{d}_{\hat{\mathbf{\Sigma}}_t} \mathbf{D}_t + d\mathbf{d}_B\,, \tag{6}$$

$$\mathbf{Z}_t = \sum_{s=1}^{t} \mathbf{d}_{A_s} \mathbf{\Sigma}^* \mathbf{d}_{A_s} + \mathbf{d}_{\mathbf{\Sigma}^*} \mathbf{D}_t + d\mathbf{d}_B\,, \tag{7}$$

where $\mathbf{d}_B = \mathrm{diag}((B_i^2)_{i\in[d]})$ and $\mathbf{d}_{\hat{\mathbf{\Sigma}}_t} = \mathrm{diag}(\hat{\mathbf{\Sigma}}_t)$ are matrices in $M_d(\mathbb{R})$.

Let $\delta \in ]0,1[$ and $f_{t,\delta} = 6d\log(\log(1+t)) + 3d\log(1+e) + \log(1/\delta)$ be an exploration parameter, we define $\mathcal{G}_t$ the event in which the estimation error remain in an ellipsoid defined by $\mathbf{Z}_t^{-1}$:

$$\mathcal{G}_t = \left\{ \left\| \sum_{s=1}^{t} \mathbf{d}_{A_s} \eta_s \right\|_{\mathbf{Z}_t^{-1}} \leq f_{t,\delta} \right\}. \tag{8}$$

Notably, this event happen at each round with high probability.

**Proposition 2.** *For events $\{\mathcal{G}_t\}_{t\leq T}$ defined as in (8), we have*

$$\textstyle\sum_{t=d(d+1)}^{T-1} \mathbb{P}\left(\mathcal{G}_t^c\right) \leq \delta T^2. \tag{9}$$

## 2.3 Algorithm

We now present `OLS-UCBV` written in Algorithm 2. The algorithm begins with an initial exploration phase by sampling every base item $i \in [d]$ and every "reachable" couple $(i,j) \in [d]^2$ at least twice. Then, for all subsequent rounds $t+1$, `OLS-UCBV` picks an action $A_{t+1}$ such that:

$$A_{t+1} \in \arg\max_{a\in\mathcal{A}} \left\{ \langle a, \hat{\mu}_t \rangle + f_t \, \left\| \mathbf{D}_t^{-1} a \right\|_{\hat{\mathbf{Z}}_t} \right\}. \tag{10}$$

**Algorithm 2** `OLS-UCBV`

**Input** $\delta > 0$, $T \geq 1$, $B \in \mathbb{R}_+^d$
**for** $t = 1, \ldots, T$ **do**
  **if** $\left\{a \in \mathcal{A} \text{ s.t } \min_{i,j \in a} n_{t,(i,j)} \leq 1\right\} \neq \emptyset$ **then**
    Choose any $A_t$ in the above set
  **else**
    Choose $A_t \in \mathcal{A}$ from (10) using $\hat{\mu}_{t-1}, \hat{\mathbf{Z}}_{t-1}$
    Environment samples $Y_t \in \mathbb{R}^d$
    Receive reward $\langle A_t, Y_t \rangle = \sum_i A_{t,i} Y_{t,i}$
    Compute $\hat{\mu}_t$ from (2)
    Compute $\hat{\mathbf{\Sigma}}_t$ from (3) and (4)
    Compute $\hat{\mathbf{Z}}_t$ from (6)
  **end if**
**end for**

### 2.4 Regret upper bound

We establish the following regret upper bounds for `OLS-UCBV`.

**Theorem 3.** *Let $T \geq 5$, $B \in \mathbb{R}_+^d$, and $\delta = 1/T^2$. Then, `OLS-UCBV` (Alg. 2) satisfies the* gap-dependent *regret upper bound*

$$\mathbb{E}[R_T] = \tilde{O}\left( \sum_{i=1}^{d} \max_{a \in \mathcal{A}/i \in a, \Delta_a > 0} \frac{\sigma_{a,i}^2}{\Delta_a} \right),$$

*and the* distribution-free *regret upper bound*

$$\mathbb{E}[R_T] = \tilde{O}\left( \sqrt{T \sum_{i=1}^{d} \max_{a \in \mathcal{A}/i \in a} \sigma_{a,i}^2} \right).$$

*where $\sigma_{a,i}^2 = \sum_{j \in a} (\mathbf{\Sigma}_{i,j}^*)_+$ for $i \in [d]$ and $a \in \mathcal{A}$, and $(\,\cdot\,)_+ = \max\{\cdot, 0\}$*

The logarithmic factors in the dimension $d$ and time $T$ are neglected in the statement. The proof rely on analyzing what happens in the events $\{\mathcal{C} \bigcap \mathcal{G}_t\}$ and is not detailed here. The thorough analysis will be available in a HAL/Arxiv version of this work.

## 3 Comparisons

We compare asymptotic regret upper bounds of different algorithms in table 1. `CUCB` for (Kveton et al., 2015) uses proxy variances and consider "worst-case" correlations, which is not really satisfactory if we know for example that some base items are independant. In order to account for this possibility, (Degenne and Perchet, 2016) introduces `OLS-UCB` which statisfies a gap-dependent regret upper bound depending on the structure. However, this algorithm needs a proxy-covariance matrix as input, which can be tricky to estimate tightly.

Perrault et al. (2020) manage to replace this by the "true" covariance matrix of the reward distribution, which is estimated online. But their upper bound also include an unsatisfactory $1/\Delta^2$ term which prevent to get a $\sqrt{T}$ gap-free upper bound. `OLS-UCB` manages to satisfies the same kind of gap-dependant upper-bound as `ESCB-C`, but to bypasses the later issue.

Another advantage of `OLS-UCBC` over `ESCB-C` is its computational complexity. The upper bounds of `OLS-UCBC` have a closed form while `ESCB-C` needs to solve linear program over convex sets to compute them.

| Algorithm | Info. | Gap-Dependant Asymptotic Regret | Gap-Free Asymptotic Regret |
|---|---|---|---|
| `CUCB` | $\Gamma$ | $d\sum_i \frac{\Gamma_{i,i}}{\min_{a/i\in a}\Delta_a}$ | $\sqrt{dT\sum_i \Gamma_{i,i}}$ |
| `OLS-UCB` | $\Gamma$ | $(1+\gamma d)\sum_i \frac{\Gamma_{i,i}}{\min_{a/i\in a}\Delta_a} + \frac{d^3 \max_i \Gamma_{i,i}}{\Delta^2_{\min}}$ | $(d^3 \max_i \Gamma_{i,i})^{1/3} T^{2/3}$ |
| `ESCB-C` | $\varnothing$ | $\frac{1}{\Delta}\sum_i \max_{a/i\in a}\sum_{j\in a}(\mathbf{\Sigma}^*_{i,j})_+ + \frac{d^3 \max_i \Gamma_{i,i}}{\Delta^2_{\min}}$ | $(d^3 \max_i \Gamma_{i,i})^{1/3} T^{2/3}$ |
| `OLS-UCBV` | $\varnothing$ | $\sum_i \max_{a/i\in a} \frac{\sum_{j\in a}(\mathbf{\Sigma}^*_{i,j})_+}{\Delta_a}$ | $\sqrt{T\sum_i \max_{a/i\in a}\sum_{j\in a}(\mathbf{\Sigma}^*_{i,j})_+}$ |

Table 1: Asymptotic $\tilde{O}(\cdot)$ regret bounds for different types of feedback, up to logarithmic terms, for the following algorithms: `CUCB` Kveton et al. (2015), `OLS-USB` Degenne and Perchet (2016), `ESCB-C` Perrault et al. (2020), and `OLS-UCBV` (ours). *Notations:* $a$ refers to actions; $i$ and $j$ refer to items; $\Gamma$ is a proxy-covariance matrix; we abbreviate $\max\{x, 0\}$ to $(x)_+$ for any $x \in \mathbb{R}$.

# References


Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Degenne, R. and Perchet, V. (2016). Combinatorial semi-bandit with known covariance. *Advances in Neural Information Processing Systems*.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. (2015). Tight regret bounds for stochastic combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Perrault, P., Valko, M., and Perchet, V. (2020). Covariance-adapting algorithm for semi-bandits with application to sparse outcomes. In *Conference on Learning Theory*.

# Analyse de données topologiques et géométriques

# Analyse Topologique de Tableaux Multiples

Rafik Abdesselam

*Laboratoires ERIC-COACTIS, Université Lumière Lyon 2, France*
*rafik.abdesselam@univ-lyon2.fr*

**Résumé.** Cet article propose une approche topologique pour explorer et analyser plusieurs tableaux de données simultanément. Il s'agit de tableaux de variables quantitatives et/ou qualitatives mesurées sur différentes thématiques homogènes, collectées sur les mêmes individus. Cette approche, appelée Analyse Topologique de Tableaux Multiples (ATTM), est basée sur la notion de graphes de voisinage dans le cadre d'une analyse conjointe de plusieurs tableaux de données. Elle permet d'étudier les éventuels liens entre plusieurs tables thématiques. La structure des corrélations ou des associations des variables de chaque tableau thématique est analysée selon le type des variables quantitatives, qualitatives ou mixtes considérées. Comme l'Analyse Factorielle Multiple (AFM), l'ATTM permet d'analyser plusieurs tableaux de variables simultanément, d'obtenir des résultats, notamment des représentations graphiques, et d'étudier la relation entre individus, variables et tableaux de données. Il peut s'agir également de tableaux de données temporelles, collectées à différents moments sur les mêmes individus. L'approche ATTM proposée est illustrée à l'aide de données réelles associées à plusieurs thématiques homogènes différentes. Les résultats sont comparés à ceux de la méthode AFM.

**Mots-clés.** Tableaux multiples, mesure de proximité, graphe de voisinage, matrice d'adjacence, analyse factorielle et classification.

**Abstract.** The paper proposes a topological approach in order to explore several data tables simultaneously. These data tables of quantitative and/or qualitative variables measured on different homogeneous themes, collected from the same individuals. This approach, called Topological Analysis of Multiple Tables (TAMT), is based on the notion of neighborhood graphs in the context of a joint analysis of several data tables. It's allows the simultaneous study of possible links between several thematic tables. The structure of the correlations or associations of the variables in each thematic table is analyzed according to the quantitative, qualitative or mixed variables considered. Like the Multiple Factorial Analysis (MFA), the TAMT allows several tables of variables to be analyzed simultaneously, and to obtain results, in particular graphical representations, which make it possible to study the relationship between individuals, variables and tables of data. These can also be tables of temporal data, collected at different times on the same individuals. The proposed TAMT approach is illustrated using real data associated with several and different homogeneous themes. Its results are compared to those from the MFA method.

**Keywords.** Multiple data tables, proximity measure, neighborhood graph, adjacency matrix, factorial analysis and clustering.

# 1 Introduction

L'objectif de cet article est de proposer une approche topologique d'analyse des données appliquée à plusieurs tableaux de données croisant les mêmes individus avec différentes variables quantitatives, qualitatives ou encore mixtes.

L'approche TAMT proposée est différente de celles qui existent déjà, notamment l'Analyse Factorielle Multiple (AMF) [Escofier et Pagès (1985), Dazy *et al.* (1996)] avec laquelle elle est comparée, ou de la méthode des Tableaux Structurants à Trois Indices de la Statistique (STATIS) [Lavit (1988)] ou encore de la méthode de l'Analyse en Composantes Doubles Principales (DPCA) [Bouroche (1975)]. Il existe désormais de nombreuses approches topologiques d'analyse factorielle et de clustering [Abdesselam (2021,2022), Aljarah *et al.* (2021), Panagopoulos (2022)] d'un seul tableau de données homogènes, mais à notre connaissance, aucune de ces approches n'a été proposée pour analyser plusieurs tableaux de données simultanément.

Le choix de la mesure de proximité parmi les nombreuses mesures existantes, joue un rôle important dans une analyse de données multidimensionnelles [Batagelj et Bren (1995), Lesot *et al.* (2009), Zighed *et al.* ( 2012)]. Elle a un fort impact sur les résultats de toute opération de structuration, de regroupement ou de clustering d'objets.

La structure de corrélation ou de dépendance des variables quantitatives ou qualitatives de chaque tableau de données dépend des données considérées. Les résultats peuvent changer selon la mesure de proximité choisie dans chaque tableau de données, qui permet de mesurer la similarité ou la dissemblance entre deux objets individus ou variables.

Ce document est organisé comme suit. Dans la section 2, nous rappelons brièvement la notion de base des graphes de voisinage, nous définissons et montrons comment construire une matrice d'adjacence associée à une mesure de proximité dans le cadre de l'analyse de la structure de corrélation ou de dépendance d'un ensemble de variables d'un tableau de données, et nous présentons les principes de l'approche proposée. Cette dernière est illustrée dans la section 3 à l'aide d'un exemple basé sur des données réelles. Les résultats sont comparés à ceux de la classification appliquée aux résultats de l'AFM. Enfin, la section 4 présente quelques remarques sur ce travail.

# 2 Topologie et tableaux de données multiples

L'analyse topologique de données est une approche basée sur le concept de graphe de voisinage. L'idée de base est assez simple, pour une mesure de proximité donnée selon le type de variables continues ou binaires et pour une structure topologique choisie, on peut faire correspondre un graphe topologique induit sur l'ensemble des objets.

L'ATTM proposée consiste à analyser simultanément plusieurs tableaux de données collectées sur les mêmes $n$ individus, à partir des différentes variables thématiques de chaque tableau de données : $X_{(n,p_x)}, Y_{(n,p_y)}, Z_{(n,p_z)}, \cdots, T_{(n,p_t)}$.

Par exemple, pour un tableau de données $X$, on considère $E_x = \{x^1, \cdots, x^j, \cdots, x^{p_x}\}$ un ensemble de $p_x$ variables quantitatives, qualitatives ou encore mixtes [Abdesselam (2021)].

On peut, au moyen d'une mesure de proximité $u$, définir une relation de voisinage, notée $V_u$, comme étant une relation binaire sur $E_x \times E_x$. De nombreuses possibilités existent pour construire cette relation de voisinage. Ainsi, pour une mesure de proximité $u$ donnée, nous pouvons construire un graphe de voisinage sur $E_x$, où les sommets sont les variables et les arêtes sont définies à partir de la propriété de la relation de voisinage.

De nombreuses définitions sont possibles pour construire cette relation binaire selon le graphe de voisinage choisi, l'Arbre de Longueur Minimalle (ALM), le Graphe de Gabriel (GG), ou encore, comme c'est le cas ici, le graphe des voisins relatifs (GVR). Pour une propriété de voisinage donnée (ALM, GG ou GVR) et une mesure de proximité $u$ choisie, parmi les nombreuses mesures existantes, on peut par exemple pour un tableau de données $X$, générer une structure topologique sur $E_x$ qui est entièrement décrite par la matrice d'adjacence binaire associée $V_{u_x}$. d'ordre $p_x$, où toutes les paires de variables voisines dans $E_x$ satisfont la propriété GVR suivante :

$$V_{u_x}(x^k\ ,\ x^l) = \begin{cases} 1 & \text{si } u(x^k\ ,\ x^l) \ \leq \ \max[u(x^k\ ,\ x^t), u(x^t\ ,\ x^l)]\ ; \\ & \qquad \forall x^k,\ x^l,\ x^t \in E,\ x^t \neq x^k \ \ et \ \ x^t \neq x^l \\ 0 & \text{sinon.} \end{cases}$$

Cela signifie que si deux variables $x^k$ et $x^l$ qui vérifient la propriété GVR sont connectées par une arête, les sommets $x^k$ et $x^l$ sont voisins.

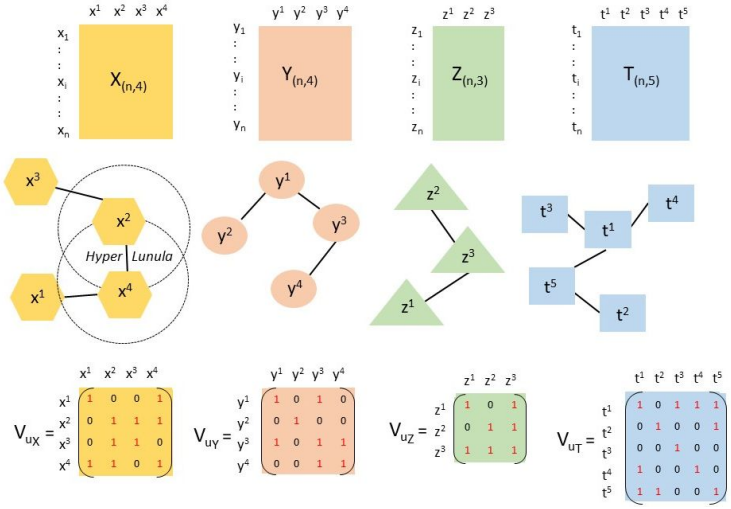



Figure 1: Tablaux multiples - Graphes et matrices d'adjacence associées

La Figure 1 montre un exemple simple dans $R^2$ de tableaux de variables de quatre thématiques observées sur les mêmes $n$ objets, selon la structure de voisinage GVR et la distance euclidienne pour chaque thématique.

Par exemple, pour le tableau de données de la thématique $X$, la valeur d'adjacence entre la deuxième et la quatrième variable, $V_{u_x}(x^2, x^4) = 1$, géométriquement, cela signifie que l'hyper-Lunule (intersection entre les deux hypersphères centrées sur les deux variables $x^2$ et $x^4$) est vide. Cela génère une structure topologique basée sur les $p_x$ objets dans $E_x$ qui sont complètement décrits par la matrice binaire d'adjacence $V_{u_x}$.

## 2.1 Matrices d'adjacence de référence

Nous analysons d'abord de manière topologique les structures de corrélation ou de dépendance des variables de chaque tableau de données, pour réaliser une analyse factorielle globale et conjointe de ces multiples tableaux, puis nous établissons sur cette analyse simultanée, une classification des individus.

Pour chaque tableau de données, $X$ par exemple, on construit la matrice d'adjacence de référence notée $V_{u_x^\star}$, soit à partir de la matrice de corrélation pour des variables quantitatives, soit à partir des profils du tableau de Burt pour des variables qualitatives. Les définitions et expressions des matrices d'adjacence de référence selon le type de variables considérées sont données dans [Abdesselam (2021), (2008)].

Pour analyser la structure de corrélation entre les variables quantitatives du tableau de données $X$ par exemple, nous examinons la signification de leur coefficient de corrélation linéaire. La matrice d'adjacence peut s'écrire comme suit en utilisant le test t de Student du coefficient de corrélation linéaire $\rho$ de Bravais-Pearson :

Pour les variables quantitatives du tableau de données $X$, la matrice d'adjacence de référence $V_{u_x^\star}$ associée à la mesure de référence $u_x^\star$ est définie comme suit :

$$\mathrm{V}_{u_x^\star}(x^k, x^l) = \begin{cases} 1 & \text{si} \quad \text{p-valeur } = P[\ |\ T_{n-2}\ | > \text{t-valeur }\ ] \leq \alpha\ ;\ \forall k, l = 1, p \\ 0 & \text{sinon.} \end{cases} \tag{1}$$

Test de Student de signification du coefficient de corrélation linéaire. L'hypothèse nulle $H_0$ d'absence de corrélation est rejetée avec une p-valeur inférieure ou égale à un risque d'erreur $\alpha$ choisi, par exemple $\alpha = 5\%$. La p-valeur est la probabilité d'accepter ou de rejeter $H_0$.

Quelle que soit le type des variables du tableau $X$, la matrice d'adjacence de référence construite $V_{u_x^\star}$ sera associée à une mesure de proximité inconnue notée $u_x^\star$. On obtient ainsi autant de matrices d'adjacence de référence que de tableaux de données considérés.

La robustesse dépend du risque d'erreur $\alpha$ choisi pour l'hypothèse nulle, pas corrélation linéaire dans le cas de variables quantitatives ou indépendance dans le cas de variables qualitatives. On peut fixer un seuil minimum afin d'analyser la sensibilité des résultats. Certes les résultats numériques seront différents, mais probablement pas leur interprétation.

## 2.2 Analyse factorielle & Classification

Pour définir l'ATTM, nous utiliserons les notations suivantes :

- On note $G_{(n,p)} = [X_{(n,p_x)}|\cdots|Y_{(n,p_y)}|\cdots|T_{(n,p_t)}]$ le tableau de données global, juxtaposition de tous les tableaux de données considérés, à $n$ lignes-individus et $p = p_x + p_y + \cdots + p_t$ colonnes-variables,

- $X_{(n,p_x)}$ est le tableau de données à $n$ individus et $p_x$ variables,

- $V_{u^\star{}_x}$ est la matrice d'adjacence symétrique d'ordre $p_x$, associée à la mesure de référence $u_x\star$ qui structure au mieux les corrélations des variables du tableau de données $X$,

- $V_{u^\star(p)} = Diag[V_{u^\star{}_x}, V_{u^\star{}_y}, \cdots, V_{u^\star{}_t}]$ est la matrice d'adjacence globale diagonale d'ordre $p$, associée à la matrice de données globale $G$,

- $\widehat{G}_{(n,p)} = GV_{u^\star}$ est la matrice des données projetées à $n$ individus et $p$ variables,

- $M_p$ est la matrice des distances d'ordre $p$ dans l'espace des individus,

- $D_n = \frac{1}{n} I_n$ est la matrice diagonale des poids d'ordre $n$ dans l'espace des variables.

**Définition 1 :** L'ATTM qui analyse simultanément les structures de corrélation de tous les tableaux de données, consiste à réaliser l'ACP standardisée du triplet $(\widehat{G}, M_p, D_n)$ [Caillez et Pagès (1976), Lebart (1989)] de la matrice de données projetée $\widehat{G} = GV_{u^\star}$.

**Définition 2 :** La classification ATTM consiste à appliquer une CAH selon le critère de Ward[1] sur les facteurs significatifs de l'analyse factorielle ATTM.

L'analyse factorielle ATTM est comparée à la méthode AFM et la classification ATTM à la méthode CAH-MFA [Fowlkes et Mallows (1983), Hubert et Arabie (1985)].

Enfin, l'approche ATTM et son dendrogramme sont facilement programmables à partir des procédures ACP et CAH des logiciels SAS, SPAD ou R.

# 3 Exemple illustratif

Panorama des régions métropolitaines de France en 2021 : pour illustrer l'approche ATTM, on a utilisé les données de l'Insee[2][Bilan économique, Rte, Inégalités, Empreinte (2021)] sur l'état des 13 régions métropolitaines de France. On a considéré quatre thématiques régionales : les Energies Renouvelables, le Climat & Environnement, le Dynamisme Economique et la Cohésion Sociale. Les libellés et satistiques sommaires des variables thématiques sont consignés dans la Table 1.

Le tableau 2 présente la matrice d'adjacence globale de référence $V_{u_\star}$ associée à la mesure de proximité $u_\star$, mesure la plus adaptée à chacun des quatre tableaux de données considérés, construite selon l'expression (1).

A noter que dans ce cas de variables quantitatives, on considère que deux variables corrélées positivement sont liées et que deux variables corrélées négativement sont également liées, mais distantes, on prendra donc en compte le signe de la corrélation des variables dans la matrice d'adjacence.

On a d'abord effectué une ATTM pour identifier la structure des corrélations des variables thématiques, puis réalisé une CAH sur les principaux facteurs de cette approche pour établir une typologie des régions selon les différentes thématiques. Les résultats de l'approche ATTM sont comparés à ceux de l'AFM.

A titre de comapaison, la Figure 2 et le Tableau 3 présentent sur le premier plan factoriel, les corrélations entre les facteurs principaux et les variables initiales. Comme on peut le

[1] Agrégation basée sur le critère de perte d'inertie minimale.

[2] Insee - Institut National de la Statistique et des Etudes Economiques

constater, ces corrélations sont légèrement différentes, tout comme les pourcentages d'inerties expliquées sur les premiers plans principaux des méthodes ATTM et AFM.

La Figure 2 montre que les deux premiers facteurs de l'ATTM expliquent respectivement 41.39% et 26.96%, représentant 68.34% de la variation totale, alors que les deux premiers facteurs de l'AFM totalisent que 59.34%.

Table 1: Dictionnaire & Statistiques sommaires des variables thématiques

| Identifier | Variable label | Identifier | Mean | Std.Deviation | Min | Max |
|---|---|---|---|---|---|---|
| | Theme: Renewable Energies (RE) | | | | | |
| CCRE | Coverage of Electricity Consumption by RE Production (%) | CCRE | 24,28 | 12,78 | 1,90 | 44,70 |
| CCWP | Coverage of Consumption of Wind power Production (%) | CCWP | 8,01 | 6,23 | 0,40 | 20,60 |
| CCSP | Coverage of Consumption of Solar Production (%) | CCSP | 3,65 | 3,32 | 0,20 | 10,80 |
| CCHP | Coverage of Consumption of Hydraulic Production (%) | CCHP | 10,98 | 12,29 | 0,00 | 39,30 |
| CCBP | Coverage of Consumption of Bioenergy Production (%) | CCBP | 1,65 | 0,65 | 0,20 | 3,00 |
| | Theme - Climate & Environment (CE) | | | | | |
| HSUN | Hours of Sunshine (h) | HSUN | 2068,46 | 292,68 | 1847,00 | 2781,00 |
| HRAI | Height of Rainfall (mm) | HRAI | 606,92 | 84,94 | 442,00 | 731,00 |
| NPSI | Number of Polluted Sites - Pollution | NPSI | 479,15 | 349,09 | 8,00 | 1131,00 |
| CARB | Carbon Footprint (tCO2e/hab) | CARB | 9,69 | 0,72 | 9,02 | 11,27 |
| CFOR | Cover Forests (%) | CFOR | 30,56 | 14,62 | 12,47 | 63,36 |
| | Theme: Economic Dynamism (ED) | | | | | |
| BCRE | Business Creation | BCRE | 20689,08 | 19151,84 | 1452,00 | 79656,00 |
| BFAI | Business Failure | BFAI | 2319,23 | 1877,07 | 142,00 | 6672,00 |
| GDPC | GDP per capita (M.€.) | GDPC | 33,62 | 8,488 | 29.12 | 62,11 |
| EMPL | Employment France (%) | EMPL | 0,08 | 0,05 | 0,01 | 0,23 |
| | Theme: Social Cohesion (SC) | | | | | |
| UNEM | Unemployment rate (%) | UNEM | 0,08 | 0,01 | 0,07 | 0,09 |
| POVE | Poverty rate (%) | POVE | 8,21 | 1,62 | 5,80 | 10,70 |
| BASI | Beneficiaries of Active Solidarity Income (RSA (%) | BASI | 0,08 | 0,05 | 0,00 | 0,21 |
| RABO | Recipients of the activity bonus (%) | RABO | 0,08 | 0,04 | 0,00 | 0,15 |
| SAEL | Social Assistance for the Elderly (%) | SAEL | 0,08 | 0,03 | 0,01 | 0,13 |
| SADP | Social Assistance for Disabled People (%) | SADP | 0,08 | 0,04 | 0,01 | 0,14 |
| CSAS | Child Social Assistance (%) | CSAS | 0,08 | 0,04 | 0,00 | 0,15 |
| MSLH | Median Standard of Living of Households (M.€.) | MSLH | 21,52 | 0,88 | 20,11 | 23,86 |

Table 2: Matrice d'adjacence globale de référence

$$V_{u^*} = \left(\begin{array}{cccccccccc|cccc|cccccccc}
1&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
\hline
0&0&0&0&0&1&-1&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&-1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
\hline
0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&0&0\\
\hline
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&1&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1
\end{array}\right)$$

Le premier plan factoriel de l'ATTM fournit une synthèse de la structure des corrélations des variables thématiques des régions françaises en 2021. Les corrélations significatives entre les variables initiales et les facteurs principaux des deux approches sont légèrement différentes. A titre de comparaison, la Figure 3 présente les dendrogrammes des classifications ATTM et AFM des régions de France selon les quatre thématiques considérées. A noter que les typologies en 4 classes de régions des deux approches sont identiques. En effet, les compositions des 4 classes sont identiques tandis que les caractérisations de ces classes sont légèrement différentes. De plus, le pourcentage de variance totale expliqué par l'approche

ATTM, $R^2 = 72.82\%$, est bien supérieur à celui de l'approche AFM, $R^2 = 66.47\%$, indiquant ainsi que les classes de l'approche ATTM sont plus homogènes que celles de l'AFM.

Table 3: Corrélations Variables & Facteurs

| **ATTM** Variable | Facteur F1 | F2 |
|---|---|---|
| CCRE | 0,025 | **-0,736** |
| CCWP | -0,227 | 0,417 |
| CCSP | -0,294 | **-0,724** |
| CCHP | 0,025 | **-0,736** |
| CCBP | 0,012 | 0,486 |
| HSUN | 0,280 | **-0,822** |
| HRAI | 0,211 | **0,854** |
| NPSI | 0,320 | **-0,716** |
| CARB | -0,101 | -0,001 |
| CFOR | -0,250 | **-0,879** |
| BCRE | **0,938** | 0,017 |
| BFAI | **0,938** | 0,017 |
| GDPC | **0,938** | 0,017 |
| EMPL | **0,938** | 0,017 |
| UNEM | 0,342 | **-0,746** |
| POVE | 0,317 | **-0,757** |
| BASI | **0,946** | 0,032 |
| RABO | **0,951** | 0,051 |
| SAEL | **0,951** | 0,051 |
| SADP | **0,951** | 0,051 |
| CSAS | **0,951** | 0,051 |
| MSLH | **0,649** | 0,282 |

| **AFM** Variable | Facteur F1 | F2 |
|---|---|---|
| CCRE | -0,276 | -0,480 |
| CCWP | 0,098 | **0,652** |
| CCSP | **-0,619** | **-0,531** |
| CCHP | -0,185 | **-0,713** |
| CCBP | 0,295 | **0,503** |
| HSUN | **-0,676** | **-0,674** |
| HRAI | 0,373 | **0,564** |
| NPSI | **0,615** | -0,019 |
| CARB | -0,080 | 0,154 |
| CFOR | **-0,545** | **-0,733** |
| BCRE | **0,792** | **-0,521** |
| BFAI | **0,683** | **-0,602** |
| GDPC | **0,677** | -0,411 |
| EMPL | **0,884** | -0,418 |
| UNEM | 0,224 | -0,361 |
| POVE | -0,069 | **-0,725** |
| BASI | **0,856** | -0,423 |
| RABO | **0,887** | -0,380 |
| SAEL | **0,759** | -0,376 |
| SADP | **0,896** | -0,332 |
| CSAS | **0,927** | -0,161 |
| MSLH | **0,619** | -0,216 |

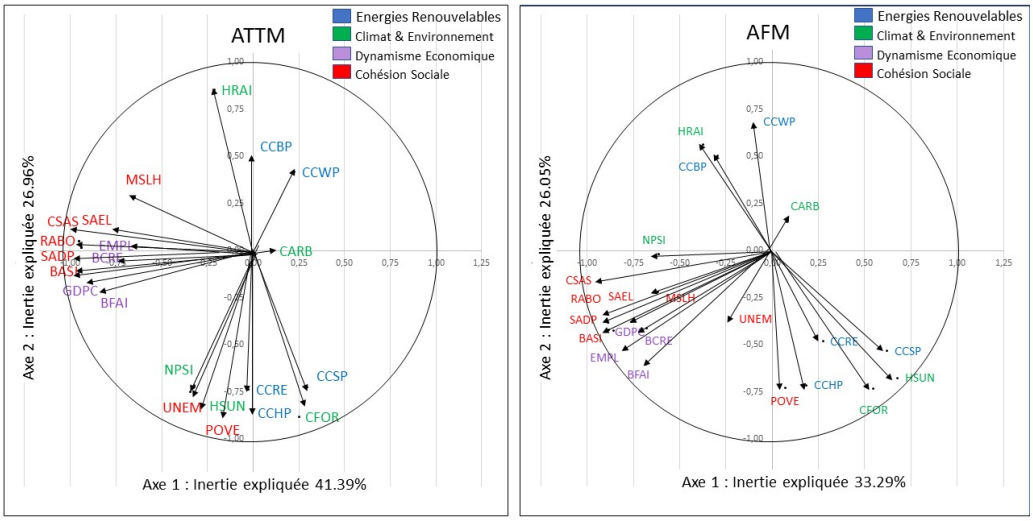



Figure 2: Représentations des variables thématiques

La figure 4 illustre en 4 couleurs, les typologies ATTM et AFM sur la carte des régions métropolitaines de France. A titre de comparaison, la Figure 4 résume les résultats significatifs des profils (+) et anti-profils (-) des deux typologies, avec un risque d'erreur inférieur ou égal à 5%. Les caractérisations diffèrent très peu, les différences sont repérées et précisées en gras et avec un astérisque.

Le premier classe ATTM, composée de six régions (Auvergne-Rhône-Alpes, Grand-Est, Occitanie, Provence-Alpes-Côte-Azur), est caractérisée par une forte couverture de la consommation électrique par les EnR, notamment par la Production Hydraulique, relativement

à la moyenne nationale des variables thématiques. Elle compte un nombre important de sites pollués nuisibles au Climat et à l'Environnement. Ces régions comptent une proportion significativement élevée de bénéficiaires du RSA, de la prime d'activité, des aides sociales aux personnes âgées, aux personnes handicapées et à l'enfance.

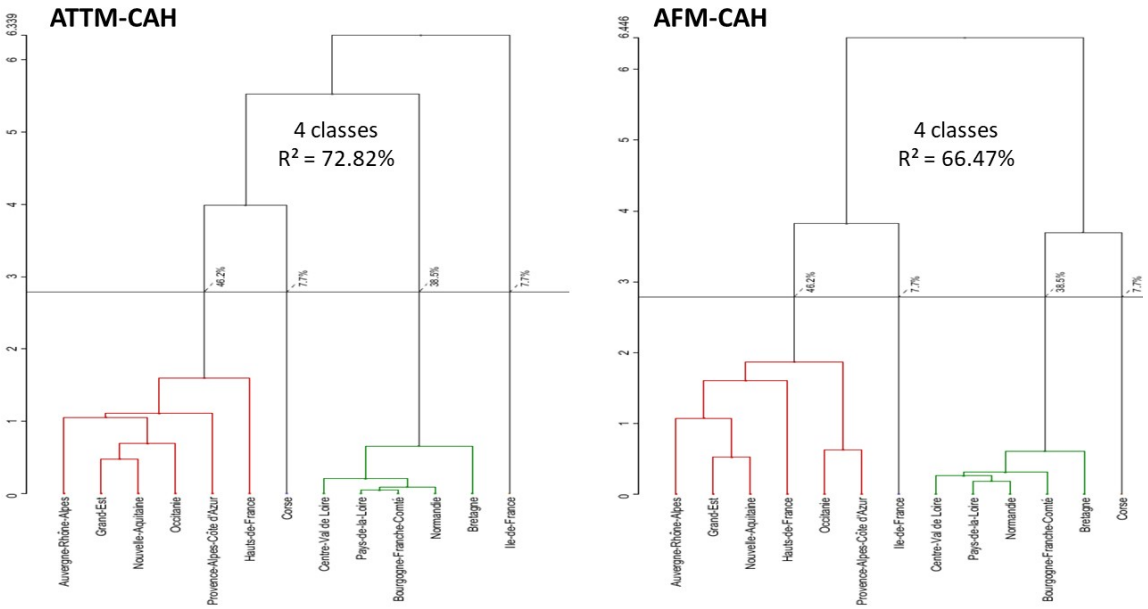



Figure 3: Arbres hiérarchiques des régions métropolitaines de France

La deuxième classe représente la région Corse, qui se caractérise par une couverture importante de la consommation d'électricité solaire et une forte couverture forestière. Elle présente également une faible couverture de la consommation d'électricité bioénergétique et de faibles précipitations d'un point de vue climatique. Cette région compte une faible proportion de bénéficiaires du RSA, de la prime d'activité, et des aides sociales aux personnes âgées, aux personnes handicapées et à l'enfance.

La troisième classe, regroupant 5 régions (Bretagne, Centre Val-de-Loire, Pays de la Loire, Bougogne-Franche-Comté et Normandie), est caractérisée par une faible couverture de la consommation électrique par les EnR et plus particulièrement par la Production Hydraulique, par rapport à la moyenne de la France métropolitaine. Les régions de cette classe présentent un faible nombre de sites pollués et un faible ensoleillement. Ces régions comptent une proportion significativement faible de bénéficiaires du RSA, de la prime d'activité, des aides sociales aux personnes âgées, aux personnes handicapées et l'enfance. Ainsi que des taux de pauvreté et de chômage faibles par rapport au niveau national.

La dernière classe représente la région Ile-de-France caractérisée par un nombre important de créations et de faillites d'entreprises, un PIB par habitant élevé et un pourcentage d'emplois élevé par rapport à la moyenne nationale. Cette région compte une proportion significativement élevée de bénéficiaires du RSA, de la prime d'activité, des aides sociales aux personnes âgées, aux personnes handicapées et à l'enfance. Le niveau de vie médian des ménages y est également très élevé.

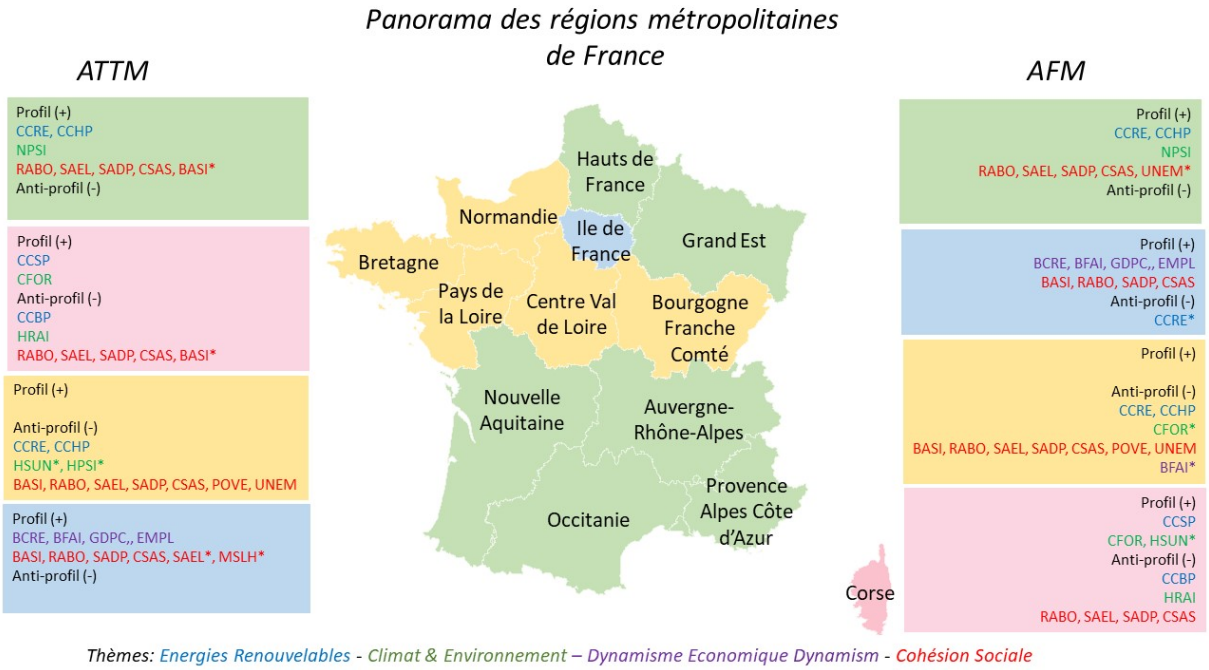



Figure 4: Typologies des classes régionales selon les thématiques

# 4 Conclusion

Cet article propose une nouvelle approche topologique pour analyser simultanément plusieurs tableaux de données, qui peut enrichir les méthodes classiques d'analyse de données. Les résultats de cette approche factorielle et clustering, basée sur la notion de graphe de voisinage, sont meilleurs que ceux de la méthode classique AFM, selon les pourcentages d'inerties expliquées par les facteurs principaux, et le $R^2$. Il serait intéressant de réaliser un Benchmark pour évaluer les résultats de cette approche topologique sur des tableaux de données massives (big data). Les travaux futurs consistent à étendre cette approche topologique à d'autres méthodes d'analyse de données, notamment dans le cadre des modèles de prédiction.

# Bibliographie

Abdesselam, R. (2022) *A Topological Clustering of Individuals*. Classification and Data Science in the Digital Age. In the Springer book series "Studies in Classification, Data Analysis, and Knowledge Organization". Edts P. Brito, J-G. Dias, B. Lausen, A. Montanari and R. Nugent, 2022.

Abdesselam, R. (2021) *A Topological Clustering of variables*. Journal of Mathematics and System Science. David Publishing Company,Vol.11, Issue 2, pp.1-17, 2021.

Abdesselam, R. (2008) *Analyse en Composantes Principales Mixte.* Classification : points de vue croisés, RNTI-C-2, Cépaduès Editions, 31-41, 2008.

Aljarah, I., Faris, H. and Mirjalali S. (2021) *Evolutionary data clustering: algorithms and applications*, Springer, 2021.

Panagopoulos, D. (2022) *Topological data analysis and clustering.* Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning, 2022.

Dazy, F., Le Barzic, J.F., Saporta, G., Lavallard F. (1996) *L'analyse des données évolutives – Méthodes et applications.* Editions TECHNIP, 1996.

Escofier, B. et Pagès, J. (1985) *Mise en oeuvre de l'AFM pour des tableaux numériques, qualitatifs, ou mixtes.* Publication interne de l'IRISA, 429, 1985.

Bouroche, J.M. (1975) *Analyse des données ternaires : la double analyse en composantes principales.* Thèse, 1975.

Lavit, C. (1988) *Analyse conjointe de tableaux quantitatifs.* Editions Masson, 1988.

Batagelj, V., Bren, M. (1995) *Comparing resemblance measures.* In Journal of classification, 12, 73–90, 1995.

Caillez, F. and Pagès, J.P. (1976) *Introduction à l'Analyse des données.* SMASH, Paris, 1976.

Lebart, L. (1989) *Stratégies du traitement des données d'enquêtes.* La Revue de MODULAD, 3, 21–29, 1989.

Lesot, M. J., Rifqi, M. and Benhadda, H. (2009) *Similarity measures for binary and numerical data: a survey.* In IJKESDP, 1, 1, 63-84, 2009.

Zighed, D., Abdesselam, R., and Hadgu, A. (2012) *Topological comparisons of proximity measures.* In the 16th PAKDD 2012 Conference. In P.-N. Tan et al., Eds. Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg, 379–391, 2012.

Bilans économiques 2021 des régions françaises.
https://www.insee.fr/fr/information/6456000

Panorama de l'électricité renouvelable 31/12/2021,
https://assets.rte-france.com/prod/public/2022-02/Pano-2021-T4.pdf.

La pauvreté dans les régions. Observatoire des inégalités, 2021.
https://www.inegalites.fr/La-pauvrete-dans-les-regions.

Carte de France de l'empreinte carbone par région, édition 2021.
https://www.hellocarbo.com/empreinte-carbone-francais-2021-par-region/.

Fowlkes, E.B., Mallows, C.L. (1983) *A Method for Comparing Two Hierarchical Clusterings.* Journal of the American Statistical Association, 78(383), 53-–569, 1983.

Hubert, L. and Arabie, P. (1985) *Comparing partitions.* Journal of Classification, 193-–218, 1985.

# Differentiable Mapper for Topological Optimization of Data Representation

Ziyad Oulhaj [1] & Mathieu Carrière [2] & Bertrand Michel [3]

[1] *Nantes Université, Centrale Nantes, Laboratoire de Mathématiques Jean Leray, France, ziyad.oulhaj@ec-nantes.fr*
[2] *DataShape, Centre Inria d'Université Côte d'Azur, France, mathieu.carriere@inria.fr*
[3] *Nantes Université, Centrale Nantes, Laboratoire de Mathématiques Jean Leray, France, bertrand.michel@ec-nantes.fr*

**Résumé.** La représentation et la visualisation non supervisées de données à l'aide d'outils de topologie constituent un domaine actif et en expansion de l'Analyse de Données Topologiques (TDA) et de la science des données. Une de ses lignes de travail les plus remarquables repose sur le graphe Mapper, qui est un graphe combinatoire dont les structures topologiques (composantes connexes, branches, boucles) correspondent à celles des données elles-mêmes. Bien que très générique et applicable, son utilisation a été jusqu'à présent entravée par l'ajustement manuel de ses nombreux paramètres, parmi lesquels un crucial est le filtre : il s'agit d'une fonction continue dont les variations sur l'ensemble de données sont l'ingrédient principal à la fois pour construire la représentation du Mapper et pour évaluer la présence et les tailles de ses structures topologiques. Cependant, il n'existe actuellement aucune méthode pour ajuster le filtre lui-même. Dans ce travail, nous nous appuyons sur un cadre d'optimisation récemment proposé incorporant la topologie pour fournir le premier schéma d'optimisation du filtre pour les graphes Mapper. Pour y parvenir, nous proposons une version plus relaxée et plus générale du graphe Mapper, dont les propriétés de convergence sont étudiées.

**Mots-clés.** Graphe Mapper, Visualisation de Données, Analyse de Données Topologiques, Homologie Persistante.

**Abstract.** Unsupervised data representation and visualization using tools from topology is an active and growing field of Topological Data Analysis (TDA) and data science. Its most prominent line of work is based on the so-called *Mapper graph*, which is a combinatorial graph whose topological structures (connected components, branches, loops) are in correspondence with those of the data itself. While highly generic and applicable, its use has been hampered so far by the manual tuning of its many parameters—among these, a crucial one is the so-called *filter*: it is a continuous function whose variations on the data set are the main ingredient for building the Mapper representation. However, there is currently no method for tuning the filter itself. In this work, we build on a recently proposed optimization framework incorporating topology to provide the first filter optimization scheme for Mapper graphs. In order to achieve this, we propose a relaxed and more general version of the Mapper graph, whose convergence properties are investigated.

**Keywords.** Mapper Graph, Data Visualization, Topological Data Analysis, Persistent Homology.

# 1 Introduction

**Mapper graphs and TDA.** The Mapper graph introduced in [1] is an essential tool of Topological Data Analysis (TDA), and has been used many times for visualization purposes on different types of data, including, but not limited to, single-cell sequencing [2, 3], neural network architectures [4, 5], or 3D meshes [6, 7]. Moreover, its ability to precisely encode (within the graph) the presence and sizes of geometric and topological structures in the data in a mathematically founded way (through the use of algebraic topology) has also proved beneficial for highlighting subpopulations of interest, which are usually detected as peculiar topological structures of significant sizes, and identifying the key features that best explain such subpopulations against the rest of the Mapper graph. This general pipeline has become a key component in, e.g., biological inference in single-cell data sets, where differentiating stem cells can usually be recovered from branching patterns in the corresponding Mapper graphs [8].

**Contributions.** Our contribution is three-fold:

- We introduce *Soft Mapper*: a generalization of the combinatorial Mapper graph in the form of a probability distribution on Mapper graphs,
- We propose a filter optimization framework adapted to a smooth Soft Mapper distribution with provable convergence guarantees,
- We implement and showcase the efficiency of Mapper filter optimization through Soft Mapper on various data sets, with public, open-source code in `TensorFlow`.

# 2 Background on Reeb and Mapper graphs

**Reeb graphs.** Mapper graphs can be understood as numerical approximations of *Reeb graphs*, that we now define. Let $X$ be a topological space and let $f : X \to \mathbb{R}$ be a continuous function called *filter function*. Let $\sim_f$ be the equivalence relation between two elements $x$ and $y$ in $X$ defined by: $x \sim_f y$ if and only if $x$ and $y$ are in the same connected component of $f^{-1}(z)$ for some $z$ in $f(X)$. The Reeb graph $\mathrm{R}_f(X)$ of $X$ is then simply defined as the quotient space $X/\sim_f$.

**Mapper graphs.** The Mapper was introduced in [1] as a discrete and computable version of the Reeb graph $\mathrm{R}_f(\mathcal{X})$. Assume that we are given a point cloud $\mathbb{X}_n = \{X_1, \ldots, X_n\} \subseteq \mathcal{X}$ with known pairwise dissimilarities, as well as a filter function $f$ computed on each point of $\mathbb{X}_n$. The Mapper graph can then be computed with the following generic version of the Mapper algorithm:

1. Cover the range of values $\mathbb{Y}_n = f(\mathbb{X}_n)$ with a set of consecutive intervals $I_1, \ldots, I_r$ that overlap, i.e., one has $I_i \cap I_{i+1} \neq \varnothing$ for all $1 \leq i \leq r-1$.

2. Apply a clustering algorithm to each pre-image $f^{-1}(I_j)$, $j \in \{1, ..., r\}$. This defines a *pullback cover* $\mathcal{C} = \{\mathcal{C}_{1,1}, \ldots, \mathcal{C}_{1,k_1}, \ldots, \mathcal{C}_{r,1}, \ldots, \mathcal{C}_{r,k_r}\}$ of $\mathbb{X}_n$.

3. The Mapper graph is defined as the *nerve* of $\mathcal{C}$. Each node $v_{j,k}$ of the Mapper graph corresponds to an element $\mathcal{C}_{j,k}$ of $\mathcal{C}$, and two nodes $v_{j,k}$ and $v_{j',k'}$ are connected by an edge if and only if $\mathcal{C}_{j,k} \cap \mathcal{C}_{j',k'} \neq \varnothing$.

# 3 Soft Mapper construction

In this section, we introduce our new construction *Soft Mapper*, which generalizes Mapper graphs and can be used for non-convex optimization. In order to do so, we first provide a general formalization of the Mapper construction that does *not* require overlapping intervals and filter functions. Then, we use this formalization to define *Soft Mapper*, which essentially consists in a distribution on regular Mapper graphs.

## 3.1 Mapper graphs built on latent cover assignments

Let $\mathbb{X}_n = \{x_1, ..., x_n\}$ be a point cloud lying in a metric space $(X, d)$ and let $r \in \mathbb{N}^\star$. For instance, $\mathbb{X}_n$ can be obtained from sampling $X^n$ according to some distribution $\mu$. Then, let Clus be a clustering algorithm on $(X, d)$, that is assumed to be fixed in the following of this work.

**Latent cover assignments.** Any binary matrix $e \in \{0, 1\}^{n \times r}$ is then called an *r-latent cover assignment* of $\mathbb{X}_n$, where $e_{i,j} = 1$ must be understood as point $x_i$ belonging to the $j$-th element of a *latent cover* of the data. For instance, in the standard version of Mapper presented in Section 2, the latent cover is obtained from a family of pre-images of intervals that cover the range of the filter function.

The procedure to compute a Mapper graph given an $r$-latent cover assignment $e \in \{0, 1\}^{n \times r}$ is straightforward: simply replace $f^{-1}(I_j)$ by $\{x_i : e_{i,j} = 1\}$ in the generic Mapper algorithm in Section 2, then derive the pullback cover using the clustering algorithm Clus, and finally compute the Mapper graph as the nerve of the pullback cover.

**Mapper function.** Let $\mathbb{K}$ be the set of simplicial complexes of dimension less or equal to 1 (i.e., graphs) and such that their sets of vertices (i.e., their $0-$skeletons) are subsets of the power set $\mathcal{P}(\mathbb{X}_n)$. We define the Mapper complex generating function as:

$$\text{MapComp}\colon \{0, 1\}^{n \times r} \longrightarrow \mathbb{K},$$

where MapComp takes a latent cover assignment as input and creates the corresponding Mapper graph.

### 3.2 Cover assignment scheme and Soft Mapper

Now, we define stochastic schemes for generating latent cover assignments, that we call *cover assignment schemes.*

**Definition 3.1.** A *cover assignment scheme* is a double indexed sequence of random variables

$$A = (A_{i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq r}}$$

such that each $A_{i,j}$ is a Bernoulli random variable conditionally to $\mathbb{X}_n$. Let $p_{i,j}(\mathbb{X}_n)$ be the parameter of the Bernoulli distribution of $(A_{i,j}|\mathbb{X}_n)$, which is thus a function of the point cloud $\mathbb{X}_n$.

**Definition 3.2.** Let $A$ be a cover assignment scheme. The *Soft Mapper* of $A$ is defined as the associated distribution of Mapper complexes, which corresponds to the push forward measure of the distribution of $A$ by the map MapComp.

# 4 Smooth relaxation of the standard cover assignment scheme

Given some $\delta > 0$, we can now define a cover assignment scheme $A_\delta$ that approximates the degenerate cover assignment scheme $A^*$ arising from the standard Mapper graph. Specifically, using the same notations as before, and denoting each element of the cover with $I_j = [a_j, b_j]$, consider, for each $j \in \{1, ..., r\}$, the function $q_j \colon X \longrightarrow [0, 1]$ defined with:

$$x \mapsto \begin{cases} 1, & \text{if } f(x) \in [a_j, b_j] \\ \exp(1 - 1/(1 - (\frac{a_i - f(x)}{\delta})^2)), & \text{if } f(x) \in [a_j - \delta, a_j] \\ \exp(1 - 1/(1 - (\frac{f(x) - b_i}{\delta})^2)), & \text{if } f(x) \in [b_j, b_j + \delta] \\ 0, & \text{otherwise} \end{cases}$$

Now, define $A_\delta = (A_{\delta,i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq r}}$ to be the random variable in $\{0, 1\}^{n\times r}$ such that for every $(i, j) \in \{1, ..., n\} \times \{1, ..., r\}$:

$$A_{\delta,i,j} \mid \mathbb{X}_n \sim \mathcal{B}(q_j(x_i)),$$

with the $A_{\delta,i,j}$'s being jointly independent conditionally to $\mathbb{X}_n$. As for the standard cover, the Bernoulli parameter $p_{i,j} = q_j(x_i)$ depends on its associated point $x_i$ and also on the chosen filter $f$.

Moreover, notice that for every $x_i \in \mathbb{X}_n$ and $j \in \{1, ..., r\}$:

$$q_j(x_i) \xrightarrow[\delta\to 0]{} \begin{cases} 1, & \text{if } f(x_i) \in I_j \\ 0, & \text{otherwise,} \end{cases}$$

and this shows that $A_\delta \xrightarrow[\delta\to 0]{\mathcal{L}} A^*$.

# 5 Topological risk of Soft Mappers

We now switch to the problem of designing filter functions automatically for Mapper graphs using Soft Mapper. To answer this ill-posed problem, we propose to look for filter functions that are optimal with respect to some topological criteria associated to their (Soft)Mapper graphs. In particular, we focus on topological losses based on *persistent homology*.

## 5.1 Topological signature for Mapper graphs

**Persistent homology.** Persistent homology is a powerful tool that allows to encode the topological information contained in a nested family of simplicial complexes, also called a *filtered simplicial complex*, see for instance [9] for a general introduction. In the context of Mapper graphs, a variation of persistent homology called *extended* persistent homology has been proved useful, as applying it on Mapper graphs produces descriptors called *extended persistence diagrams*. These diagrams only require to define a *filtration function* on the graph, and are made of points in the Euclidean plane, each point encoding the presence and size (w.r.t. the filtration function) of a particular topological structure of the Mapper graph (such as a connected component, a branch or a loop).

We now define a filtration function on Mapper graphs in order to compute extended persistence diagrams. Let $\mathcal{F}(\mathbb{X}_n, \mathbb{R})$ be the space of real valued functions defined on the point cloud $\mathbb{X}_n$. For a function $F \in \mathcal{F}(\mathbb{X}_n, \mathbb{R})$, we first associate a filtration $\phi$ to some $K \in \mathrm{im}(\mathrm{MapComp})$ with:

$$\forall \sigma \in K \, : \, \phi(\sigma) = \max_{c \in \sigma} \frac{\sum_{x \in c} F(x)}{\mathrm{card}(c)},$$

that is, node filtration values are defined as the average filter values of the data points associated to the node, and edge filtration values are computed as the maximum of their node values. Then, we compute the extended persistence diagram (which we consider as a subset of $\mathbb{R}^2$ ) of the filtered simplicial complex $(K, \phi)$. We denote by MapPers the function that takes a Mapper graph and a scalar function on $\mathbb{X}_n$, and then outputs the persistence diagram:

$$\mathrm{MapPers}\colon \mathbb{K} \times \mathcal{F}(\mathbb{X}_n, \mathbb{R}) \longrightarrow \mathcal{P}(\mathbb{R}^2).$$

**Persistence specific loss.** Now, we introduce a generic notation for a loss function—or, more simply, a statistic—that associates a real value to any extended persistence diagram. Denoting $PD$ as the set of subsets of $\mathbb{R}^2$ consisting of a finite number of points outside the diagonal $\Delta = \{(x, x) : x \in \mathbb{R}\}$, such a function can be written as $\ell\colon PD \longrightarrow \mathbb{R}$.

## 5.2 Statistical risk of the topological signature associated to Soft Mapper

We finally compute the loss associated to a Mapper graph with the function

$$\begin{aligned}\mathcal{L}\colon \{0,1\}^{n\times r} \times \mathcal{F}(\mathbb{X}_n, \mathbb{R}) &\longrightarrow \mathbb{R} \\ (e, F) &\longmapsto \ell\left(\text{MapPers}\left(\text{MapComp}(e), F\right)\right).\end{aligned}$$

Then, we define the risk of a Soft Mapper MapComp$(A)$ by integrating the loss according to the distribution of the Soft Mapper, or equivalently according to the distribution of the cover assignment scheme:

$$\mathbb{E}\left(\mathcal{L}(A, F)|\mathbb{X}_n\right) = \sum_{e\in\{0,1\}^{n\times r}} \mathcal{L}(e, F) \cdot \mathbb{P}(A = e|\mathbb{X}_n).$$

# 6 Conditional risk optimization with respect to parameters

Now that we have properly defined risks associated to Soft Mapper distributions, we study in this section the convergence properties of filter optimization schemes minimizing such risks.

## 6.1 Problem setting

Let us introduce a parameterized family of functions $\{f_\theta : \mathbb{X}_n \to \mathbb{R},\, \theta \in \mathbb{R}^s\}$. In order to simplify notations, we assume in the following of the article that the function $F$ used to compute persistence diagrams and the filter function $f_\theta$ used to design cover assignments are the same, $F = f_\theta$. Let $A$ be a cover assignment scheme whose joint distribution $\mathbb{P}_\theta$ depends on the filter function $f_\theta$.

Our goal is to find the optimal set of parameters $\bar{\theta}$ that minimizes the topological risk associated to MapComp$(A)$, when $f_\theta$ is used to define the filtration values on the Mapper graphs. In other words, if we denote:

$$\begin{aligned}\mathrm{L}\colon \mathbb{R}^s &\longrightarrow \mathbb{R} \\ \theta &\longmapsto \mathbb{E}_\theta(\mathcal{L}(A, f_\theta)|\mathbb{X}_n),\end{aligned} \tag{1}$$

our aim is to find a minimizer of L. Note that in the definition of L, the expectation depends on $\theta$ because the distribution of $A$ also depends on it.

In order to prove guarantees about minimizing L, we follow [10], which uses the theoretical background introduced in [11], in which the authors prove that stochastic gradient descent algorithms converge under certain conditions. To use this framework, it suffices to prove two points (see Corollary 5.9. in [11]):

- L is definable in an o-minimal structure (see [12]),
- L is locally Lipschitz.

## 6.2 Theoretical guarantees on the convergence of a gradient descent scheme

Under regularity assumptions on the parameterized family of filter functions $\mathcal{F} = \{f_\theta \colon \mathbb{X}_n \longrightarrow \mathbb{R},\, \theta \in \mathbb{R}^s\}$, we now show that the risk L in Equation (1) is definable and smooth.

**Theorem 6.1.** *Suppose that there exists an o-minimal structure $\mathcal{S}$ such that:*

- *for every $x \in \mathbb{X}_n$, the function $\theta \mapsto f_\theta(x)$ is definable in $\mathcal{S}$ and is locally Lipschitz,*
- *for every $m \in \mathbb{N}$, the restriction of $\ell$ to the set of (extended) persistence diagrams of size $m$ is definable in $\mathcal{S}$ and is locally Lipschitz,*
- *for every $e \in \{0,1\}^{n \times r}$, the function $\theta \mapsto \mathbb{P}_\theta(A = e|\mathbb{X}_n)$ is definable in $\mathcal{S}$ and is locally Lipschitz.*

*Then* L *is definable in $\mathcal{S}$ and is locally Lipschitz.*

Under the assumptions of Theorem 6.1, it is then possible to give guarantees on the convergence of a stochastic gradient descent scheme to some critical points of L. This only requires additional, but mild and not very restrictive technical conditions regarding the stochastic gradient descent algorithm itself.

# 7 Numerical Experiments

A first application where we can use the Soft Mapper optimization setting is finding linear filters in order to skeletonize 3-dimensional shapes with Mapper graphs. Here, our dataset $\mathbb{X}_n$ consists each time of a point cloud embedded in $\mathbb{R}^3$. The different point clouds we study are displayed (as meshes) in Figure 1. The parametric family of functions is linear, i.e., equal to $\{f_\theta \colon x \mapsto \langle x, \theta \rangle,\, \theta \in \mathbb{R}^3\}$, and the cover assignment scheme $A_\delta$ is the smooth relaxation of the standard case, with $\delta = 10^{-2} \cdot (\max_{x \in \mathbb{X}_n} f_\theta(x) - \min_{x \in \mathbb{X}_n} f_\theta(x))$. The cover of the image space is given by $r$ intervals of the same length, such that consecutive intervals have a percentage $g$ of their length in common. The clustering algorithm for the three shapes is KMeans.

**Objective.** Intuitively, the optimal directions to filter the 3-dimensional shapes (in a topological sense) are:

- for the human: the vertical direction,
- for the octopus: the parallel direction to its tentacles,
- for the table: the perpendicular direction to its upper surface,

as these directions induce Mapper graphs with more topological structures. We will therefore measure the quality of our results by comparing our optimized directions $\bar{\theta}$ to the ones cited above. To find $\bar{\theta}$, we use the total (regular) persistence as a persistence specific loss $\ell$, each time taking the diagonal as the initial direction, i.e. $\theta_0 = (\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})^T$. The learning curves for each 3-dimensional shape are displayed in Figure 2, and the correlations between the optimized directions and those we identified as intuitively optimal are summarized in Table 1. As one can see from the table, we are able to recover these intuitive directions with gradient descent.

**Qualitative assessment.** One can see, in Figures 3 and 4, that the regular Mapper graphs built with the initial and final (optimized) filter functions show clear improvement in the ability of the graphs to act as skeletons of the original point clouds. The third shape, representing a table, is particularly interesting. Indeed, the optimal direction that we captured is different from the first and the second principal components computed by PCA, since the principal plane of the point cloud is given by the surface of the table.

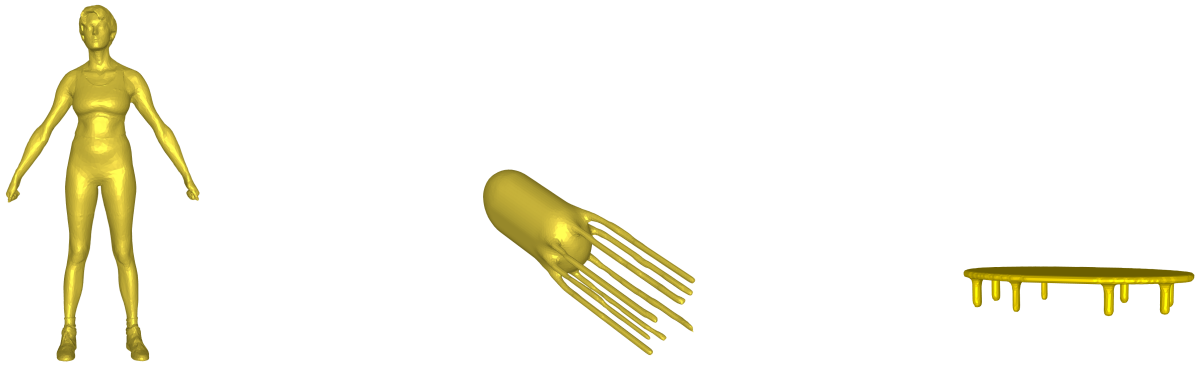


Figure 1: Meshes of 3-dimensional point clouds representing from left to right: a human, an octopus and a table.

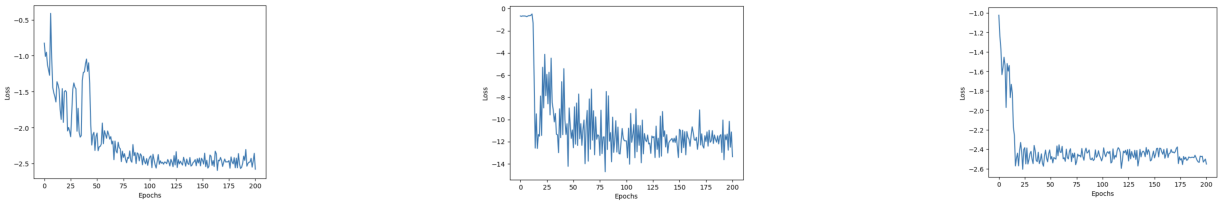



Figure 2: Learning curves for the 3-dimensional shapes corresponding, from left to right, to: the human, the octopus and the table.

| Human | Octopus | Table |
|---|---|---|
| 0.9999 | -0.9984 | 0.9993 |

Table 1: Correlation between the optimized directions and the optimal ones, for each 3-dimensional shape.

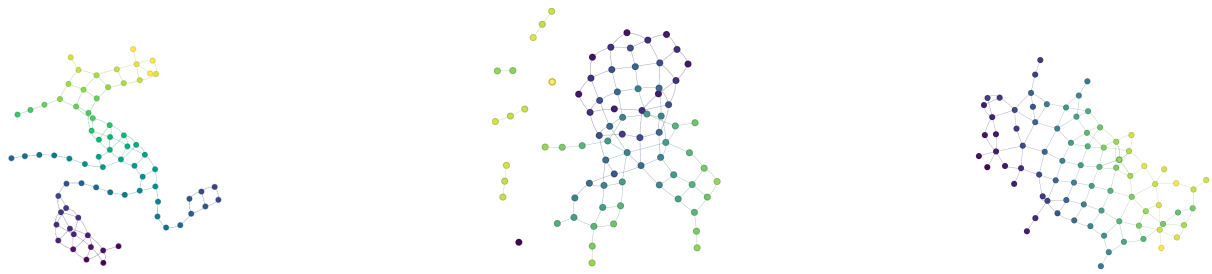


Figure 3: Regular Mapper graphs computed with the initial filter function, corresponding, from left to right, to: the human, the octopus and the table. Vertices are colored using the mean value of the filter function in the corresponding clusters.

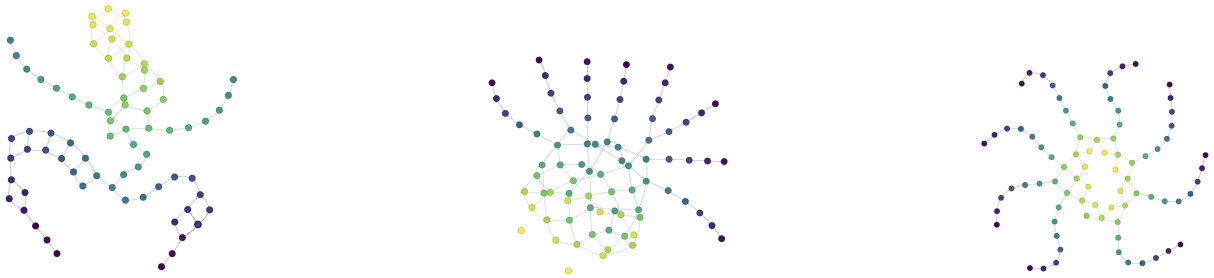


Figure 4: Regular Mapper graphs computed with the optimized filter function, corresponding, from left to right, to: the human, the octopus and the table. Vertices are colored using the mean value of the filter function in the corresponding clusters.

# 8 Discussion

In this article, we have introduced Soft Mapper, a distributional and smoother version of the standard Mapper graph, with provable convergence guarantees using persistence-based losses and risks. Our case study in this article was finding an optimal filter function, among a parameterized family of functions, in order to construct regular Mapper graphs incorporating an optimized and maximal amount of topological information. We then produced examples of such optimization processes on real 3D shape data, for which we were able to obtain structured Mapper representations in an unsupervised way.

# Bibliographie

# References

[1] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2:091–100, 2007.

[2] Tongxin Wang, Travis Johnson, Jie Zhang, and Kun Huang. Topological methods for visualization and analysis of high dimensional single-cell rna sequencing data. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 350–361. World Scientific, 2018.

[3] Sabrina Zechel, Pawel Zajac, Peter Lönnerberg, Carlos F Ibáñez, and Sten Linnarsson. Topographical transcriptome mapping of the mouse medial ganglionic eminence by spatially resolved rna-seq. *Genome biology*, 15:1–12, 2014.

[4] Satanik Mitra and Kameshwar Rao JV. Experiments on fraud detection use case with qml and tda mapper. In *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 471–472, 2021.

[5] Brian B Joseph, Trami Pham, and Christopher Hastings. Topological data analysis in conjunction with traditional machine learning techniques to predict future mdap pm ratings. Acquisition Research Program, 2021.

[6] Ziqi Wang. Exploration of topological data analysis in 3d printing. In *2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, pages 150–153. IEEE, 2020.

[7] Paul Rosen, Mustafa Hajij, Junyi Tu, Tanvirul Arafin, and Les Piegl. Inferring quality in point cloud-based 3d printed objects using topological data analysis. *arXiv preprint arXiv:1807.02921*, 2018.

[8] Abbas Rizvi, Pablo Cámara, Elena Kandror, Thomas Roberts, Ira Schieren, Tom Maniatis, and Raúl Rabadán. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35:551–560, 2017.

[9] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Society, 2010.

[10] Mathieu Carriere, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, and Yuhei Umeda. Optimizing persistent homology based functions. In *International conference on machine learning*, pages 1294–1303. PMLR, 2021.

[11] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[12] Michel Coste. *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000.

# Time topological analysis of EEG using signature theory

Stéphane Chrétien [1] & Ben Gao [2] & Rémi Vaucher [3] & Astrid Thébault Guiochon [4]

[1] *Laboratoire ERIC, Université Lyon 2, France, stephane.chretien@univ-lyon2.fr*
[2] *Halias Technologie, Grenoble, France, ben.gao@halias.fr*
[3] *Halias Technologies & Laboratoire ERIC, université Lyon 2, remi.vaucher@halias.fr*
[4] *Laboratoire EMC, Université Lyon 2, a.thebaultguiochon@univ-lyon2.fr*

**Résumé.** La détection d'anomalies dans les signaux multivariés est une tâche de première importance dans de nombreuses disciplines (épidémiologie, finance, sciences cognitives et neuro-sciences, cancérologie etc). Dans cette optique, l'analyse topologique des données (TDA) offre une batterie d'invariants "de forme" qui peuvent être exploités pour la mise en ouvre d'un schéma de détection efficace. Notre contribution consiste à étendre les construction présentées dans [5] sur la construction de complexes simpliciaux à partir des Signatures des signaux et leur capacités prédictives, plutôt que de l'utilisation d'une distance générique comme dans [12]. La théorie des Signatures est une nouvelle thématique en Machine Learning [3] issue des travaux récents sur les notions de Chemins Rugueux (Rough Paths) développés par Terry Lyons et son équipe [9] sur la base du formalisme introduit par Chen [2]. Nous explorons en particulier la détection des changements de topologie, sur la base du suivi de l'évolution de la persistance homologique et des nombres de Betti associés au complexe introduit dans [5]. Nous appliquons nos outils pour l'analyse de signaux cérébraux de type EEG afin de détecter des phénomènes précurseurs aux crises d'épilepsies.

**Mots-clés.** Analyse de données topologique, Séries Temporelles, Statistiques Appliquées

**Abstract.** Anomaly detection in multivariate signals is a task of paramount importance in many disciplines (epidemiology, finance, cognitive sciences and neurosciences, oncology, etc.). In this perspective, Topological Data Analysis (TDA) offers a battery of "shape" invariants that can be exploited for the implementation of an effective detection scheme. Our contribution consists of extending the constructions presented in [5] on the construction of simplicial complexes from the Signatures of signals and their predictive capacities, rather than the use of a generic distance as in [12]. Signature theory is a new theme in Machine Learning [3] stemming from recent work on the notions of Rough Paths developed by Terry Lyons and his team [9] based on the formalism introduced by Chen [2]. We explore in particular the detection of changes in topology, based on tracking the evolution of homological persistence and the Betti numbers associated with the complex introduced in [5]. We apply our tools for the analysis of brain signals such as EEG to detect precursor phenomena to epileptic seizures.

**Keywords.** Topological and Geometric Data Analysis, Time Series, Applied Statistics

# 1 Introduction

Topological data analysis [1] is a rapidly growing field that has emerged recently, based on the intriguing observation that data come with shape-like properties. In general, existing topological structures that are built from data, such as Cech or Vietoris Rips complexes, make essential use of a metric that may not be fully suitable because inherited from the space where our data are imbedded instead of the geodesic distance on the manifold where our data truly live. In a recent paper [5], we proposed a novel way of buiding simplicial complexes based on performance in prediction rather than distance. The main tool for performing prediction is the signature transform [3], that easily extract meaningful features from multidimensional signals. In this setting, each node represents a component of the signal and a node belongs to a simplex if the signature of all the nodes from the simplex accurately explains the considered node, statistically speaking. Similarly, a face belongs to a simplex if the signature of the signals represented by the face is accurately predicted by the adjacent faces of various orders in the simplex. In this approach, Signatures faithfully represent faces and simplices together with a natural orientation and make the topological construction better motivated and more statistically meaningful. From a computational viewpoint, selecting faces that explain another one using their respective signatures can be done efficiently using the LASSO, in the spirit of the celebrated method proposed in [10] for estimating graphical models. More recently, [12] solved this problem more precisely by considering a signal correlation matrix. This method allows to create high-dimensionnal structures.

In the present paper, we examine a more refined aspect of our topological construct : the dynamic evolution of the topology as a function of time as complexes undergo potential structural transformations at specific change points in time, reflecting the apperance of certain phenomena. In the area of neuroscience, this approach will be instrumental for detecting change points at which electroencephalograms reflects known "neuroscientific" behaviors.

## 1.1 Signature of rough paths (in a nutshell)

Consider a $d$-dimensionnal path $X = (X^1, ..., X^d) : \mathbb{R} \to \mathbb{R}^2$. In the following, we will note $S_I^{(k)}(X)$ the $k$-th degree signature applied to $X$ on an time interval $I = [a, b]$. The **signature** of $X$ is given by the tensors sequence $S_I^{(k)}(X) \in \mathbb{R}^{\underbrace{d \times d \times \cdots \times d}_{k \text{ times}}}$ for all $k \in \mathbb{N}$. We will use the notion of truncated signature of order $K$ to define the sequence of signature tensor $S_I^{(k)}, k \leq K$. These tensors are given, for $\{i_1, ..., i_k\} \subset \{1, ..., k\}$ by :

$$\left(S_I^{(k)}(X)\right)_{i_1,\dots,i_k} = S_I^{i_1,\dots,i_k}(X) = \int_{a<t_1<t_2<\cdots<t_k<b} \cdots \int dX_{t_1}^{i_1} dX_{t_2}^{i_2} \dots dX_{t_k}^{i_k}$$

The signature of a rough path serves as a powerful geometric feature extractor. To compute the signature of a multidimensional signal that is discretely sampled, we must consider linear interpolation between each consecutive time point, thus invoking Chen's theorem :

**Theorem 1.1** (Chen's identity)**.** *Consider $X : [a, b] \to \mathbb{R}^d$ and $Y : [b, c] \to \mathbb{R}^d$. Define :*

$$X * Y = \begin{cases} X_t, & si\ t \in [a, b] \\ Y_t - Y_0 + X_b, & si\ t \in [b, c] \end{cases}$$

*as the* ***concatenation*** *of $X$ and $Y$. Then :*

$$S_I^{(k)}(X * Y) = S_I^{(k)}(X) \otimes S_I^{(k)}(Y)$$

To center/rescale the signals, one has to consider the following properties

— For any constant $\gamma \in \mathbb{R}^d$, $S_I^{(k)}(X + \gamma) = S_I^{(k)}(X)$

— For any constant $\lambda \in \mathbb{R}$, $S_i^{(k)}(\lambda X) = \lambda^k S_I^{(k)}(X)$

Furthermore, to ensure a "pseudo" unicity (because it is computationally infeasible to create the whole signature) we will consider the next result :

**Theorem 1.2.** *If $\exists i \in \{1, ..., k\}$ such that $X^i$ strictly monotonic, then $S_I(X)$ uniquely defines $X$.*

Finally, defining $\mathcal{C}^{1-var}_{\text{mon}}(I, \mathbb{R}^d)$ the space of $d$-dimensionnal signals (linearly interpolated on a subdivision $D(I)$ of $I$) with $|X|_{1-var} = \sup_{t_i \in D(I)} \sum_i d(X_{t_i}, X_{t_{i+1}}) < \infty$ and such that $\exists i \in \{1, \dots, d\}$ with $X^i$ strictly monotonic, then the signature transform $S_I : \mathcal{C}^{1-var}_{\text{mon}}(I, \mathbb{R}^d) \to S_I(\mathcal{C}^{1-var}_{\text{mon}}(I, \mathbb{R}^d))$ is an homeomorphism. [7]

## 1.2 Simplicial complex

Consider a set of $d$ vertices $V = \{X^1, ..., X^d\}$.

**Definition 1.1.**
— *For $k < d$, a $k$-simplex $\sigma_k$ is a $n + 1$ set and all its subsets [14].*
— *A geometric realization of $\sigma_k$ is given by the convex hull $E_k$ of its $n+1$ points (eventually embedded) in $\mathbb{R}^d$ such that $\dim(E_k) = k$ (so $d \geq k$)*

With this tool, one can build a new structure on $V$[1].

**Definition 1.2.** *An abstract simplicial complex $\mathcal{C}$ on a finite vertex set $V$ is a collection $\{\sigma_k, k < d\}$ of simplices such that :*

— *$v \in \mathcal{C}$ if $v \in V$*

— *$\tau \subseteq \sigma \in \mathcal{C} \Rightarrow \tau \in \mathcal{C}$*

*The* ***dimension*** *of $\mathcal{C}$ is the highest $k$ such there exists a $k$-simplex in $\mathcal{C}$.*

A simplex $\sigma \in \mathcal{C}$ is called a **face** of $\mathcal{C}$. The upcoming definition is the most important for our algorithm :

**Definition 1.3.** *Consider $k \in N^*$ and $\sigma$ a $k$-simplex in a simplicial complex $\mathcal{C}$. Its* ***link*** *$Lk(\sigma, \mathcal{C})$ in $\mathcal{C}$ is the set of all faces $\tau \subset \mathcal{C}$ such that :*

— $\sigma \cap \tau = \varnothing$
— $\sigma \cup \tau$ *is a face of* $\mathcal{C}$

**Remarks :**
— The link of a vertex is called **neighborhood** in graph theory.
— In the following algorithm, for a vertex $v$ and a fixed $k$, we build the subset of all the $k$-simplex in $\mathrm{Lk}(v, \mathcal{C})$ that we call $k$-dimensional Link (for simplicity).

## 1.3 Some tools for simplicial complex analysis.

The main goal of this section is to introduce some useful topological invariants, the first of these being the **Betti numbers**[1]. Informally, each $b_k$ count the number of $k+1$-dimensional holes in $\mathcal{C}$. In our experiments, we focus on 2-dimensional complexes, and thus only consider $b_0$ (the number of connected components) and $b_1$ (the number of 2-dimensional holes).

The second invariant that we used is the **Persistence Diagram**. To get a numerical summary of this, we compute the notion of **Persistence Entropy**[13, 4]. A more robust notion of Persistence Summary presented in [6] could be put to work, but the implementation raises much more complex issues that we will not address in the present work.

**Definition 1.4.** *Consider a $k$-dimensional simplicial complex $\mathcal{C}$. A **filtration** over $\mathcal{C}$ is a sequence $(\mathcal{C}_i)_{0 \leq i \leq n}$ of simplicial complexes such that :*
— *For all $1 \leq i \leq j \leq n$, $\mathcal{C} \subset \mathcal{C}_j$*
— $\mathcal{C}_n = \mathcal{C}$

A natural way of building a filtration is by defining an increasing (non necessary strictly) sequence of time of arrival $(b_i)_{1 \leq i \leq n}$ for each simplex in $\mathcal{C}$. In Cech complex, the time of arrival correspond to the radius $r$ at wich the simplex appear in $\mathcal{C}$.

**Definition 1.5.** *Consider a simplicial complex $\mathcal{C}$ and his persistence diagram $DP$*

$$DP = \{[b_i, d_i), i \in \{1, ..., \#\mathcal{C}\}\}$$

*with $b_i$ and $d_i$ respectively birth and death times for sub-simplex of $\mathcal{C}$.*
*Define*

$$p_i = \frac{b_i - d_i}{\sum\limits_j b_j - d_j}$$

***Persistence entropy** (PE) is given by the Shannon entropy of $\{pi\}$ :*

$$PE = -\sum_i p_i \log(p_i)$$

Simplistically , the persistence entropy measures the similarity (or dissimilarity) between closure speed of $k$-dimensional holes : high level PE indicates that all holes are filled at same speed, in opposition to a PE close to 0. Since a quickly filled hole is likely just noise, PE quantifies how many significatives sub-structures lies in $\mathcal{C}$.

Beginning with section 2, we will consider these structures and invariants as evolving as a function of time : for any timestamp $t \in I$, we will consider the simplicial complex $\mathcal{C}(t)$, its associated betti numbers $b_0(t), ..., b_n(t)$ and its related persistence entropy denoted by $PE(t)$.

**Remark :** many other features may be extracted from the filtration we introduce for $\mathcal{C}$. More will be studied in a long version of this contribution.

## 1.4 Our simplicial complex construction algorithm

### 1.4.1 The algorithm

Our algorithm build the simplicial complex on $X = \{X^1, ..., X^d\}$ by constructing the $k$-dimensional link (iteratively on $k$) of a channel $X^i$. Here is a simplified version of the algorithm [5].

**Data:** Fix the degree *deg* of signature and a computation interval $I$.
**Result:** The simplicial complex of interaction among each time series.
**for** *k from 1 to* $d-1$ **do**
  **for** *i from 1 to d* **do**
    Compute $S^{deg}(\tilde{X^i})$ (where $\tilde{X^i}(t) = (t, X^i(t))$).;
    For every word $i_1...i_k$ of $\{1, ..., d\} \setminus \{i\}$, compute $S_I^{deg}(\tilde{X}^{i_1...i_k})$ where $\tilde{X}^{i_1...i_k}(t) = (t, X^{i_1}, ..., X^{i_k})$;
    Predict $S_I^{deg}(\tilde{X^i})$ from $\left\{S_I^{deg}(\tilde{X}_{i_1...i_k})\right\}_{i_1...i_k \in (\{1,...,d\}\setminus\{i\})^k}$ with LASSO;
    **if** $R^2 > 0,67$ **then**
      Select all non-zero $\beta_{i_1...i_k}$
    **else**
      Fix $\beta_{i_1...i_k} = 0$ for all $i_1...i_k$.
    **end**
    **for** *Every* $\beta_{i_1...i_k} \neq 0$ **do**
      Create the simplex whose vertices are $\{X^i, X^{i_1}, ..., X^{i_k}\}$
    **end**
  **end**
**end**

**Algorithm 1:** Sequential construction of a simplicial complex on $(X^1, ..., X^d)$

### 1.4.2 The induced filtration :

In order to retrieve the persistence diagram, we need to construct a filtration and so birth time.

— Take all the weights $(\beta_\sigma)_{\sigma\in\mathcal{C}}$ attached to each simplex.
— Create $b_\sigma = 1 - \frac{\beta_\sigma}{\sum_{\sigma\in\mathcal{C}}}\beta_\sigma$.
— For each $\tau, \sigma \in \mathcal{C}$ such that $\tau \subset \sigma$ and $b_\tau > b_\sigma$, then fix $b_\tau = b_\sigma$.

This (non necessarily strictly) increasing sequence ensures that a highly significative simplex appears early in the filtration. In the following section, we show how to exploit this filtration.

# 2 Empirical study

We performed a set of computational experiements using EEG signals from the CHB-MIT Scalp Database [8]. The main benefit of using real-world signals is to test the stability of our algorithm against a uncontrolled noise.

## 2.1 Importance of hyperparameter values

We will evaluate the impact of the choice of the main parameters based on a one-hour trajectory, averaged every second. The signals reflect the occurence of an epilepsy seizure which t lasts less than one minute (on the interval $[49', 51']$). We focus our exploration on the LASSO parameters and the sliding Window size for the computation of the Signature. For the sake of simplicity, we will specify the maximal dimension of $\mathcal{C}(t)$ as equal to 2.

**LASSO parameters** For every channel $X^i$, we solve the following minimization problem :

$$\min_{\beta\in\mathbb{R}} \|S_I^{deg}(\tilde{X}^i) - \sum_{j\neq i} \beta_j S_I^{deg}(\tilde{X}^j)\|_2^2 + \lambda_1 \sum_{j\neq i} |\beta_j| \tag{1}$$

LASSO allows us to select which signatures (and then by homeomorphism which $X^j$) lies in the neighbourhood of $X^i$. This selection depends entirely on $\lambda_1$. However, this minimization only creates the 1-dimensional neighbourhood of $X^i$. to creates the 2-dimensional NH, we solve :

$$\min_{\beta\in\mathbb{R}} \|S_I^{deg}(\tilde{X}^i) - \sum_{j_1,j_2\neq i, j_1\neq j_2 i} \beta_{j_1 j_2} S_I^{deg}(\tilde{X}^{j_1 j_2})\|_2^2 + \lambda_2 \sum_{j\neq i} |\beta_{j_1 j_2}| \tag{2}$$

This naturally brings a second parameter.

Experimental results included in the version posted on ArXiv (more precisely, see Figure 3 and 5 from the appendix of the ArXiv version), show that fixing $\lambda_1$ and varying $\lambda_2$

does not change $b_1$ nor the $PE$ trajectory. Then, fixing $\lambda_2$ and varying $\lambda_1$ does not give the same result : on the one hand, similarly to the precedent case, persistence entropy is not much impacted by how $\lambda_1$ varies. On the other hand, $b_1$ 's trajectory becomes noisier when one enforces sparsity in the LASSO.

**Sliding window size for computing the Signature :** At each time $t$ we build a simplicial complex $\mathcal{C}_t$, and for a path $P$ defined on $I$, we compute $S^k_{[t-L,t]}(P)$ (under the condition $[t-L,t] \subset I$). Length $L$ of $I$ is then a hyperparameter that can impact the topological structure of $\mathcal{C}(t)$. First results show that choosing $L$ is key to safe detection of the sought for patterns.

## 2.2 Experiments with multiple EEG trajectories.

In the proposed experiments, we focus on a multivariate EEG signal sampled during 15 consecutives hours, and that includes 3 seizures. The method's hyperparameters are specified as follow :

— $\lambda_1 = 1$ and $\lambda_2 = 1$

— $L = 50$.

The main goal is to retrieve pre-critical and/or critical behaviour based on the Betti numbers and the Persistence Entropy, our main topological invariants of interests. According to [11],

**Pre-critical behaviour :** are caracterized by "area partitioning" (loss of connectivity between neuronal areas, loss of synchrony). Our main hope is that this behaviour is reflected in the trajectory of $b_1$ : a loss of connectivity should result in poorly interconnected cortical areas, resulting in several stabilized 2-dimensional holes. An expected dynamics is then to observe the steady growth of $b_1$ in time. It would be relevant to extend the dimension to $k = 3$ in order to see if this fact is confirmed with $b_2(t)$. Another idea would be to caracterise every hole in $\mathcal{C}$ and track the stable ones among them.

**Critical behaviour :** Neuronal populations will abruptly synchronise throughout the duration of the crisis, leading to hypersynchrony before returning to a more 'local' level of synchrony.

In Figure 1 below, one observes the global behavior of $b_1$ and $PE$ on 1 hour of time (second trajectory). We computed an average measure on a clean trajectory (the green line) for both $b_1$ and $PE$. The grey area represent the interval $[x(t) - \hat{\sigma_x}, x(t) + \hat{\sigma_x}$ with $x = b_1$ or $PE$ and $\hat{\sigma_x}$ computed on $[t-h,t]$.

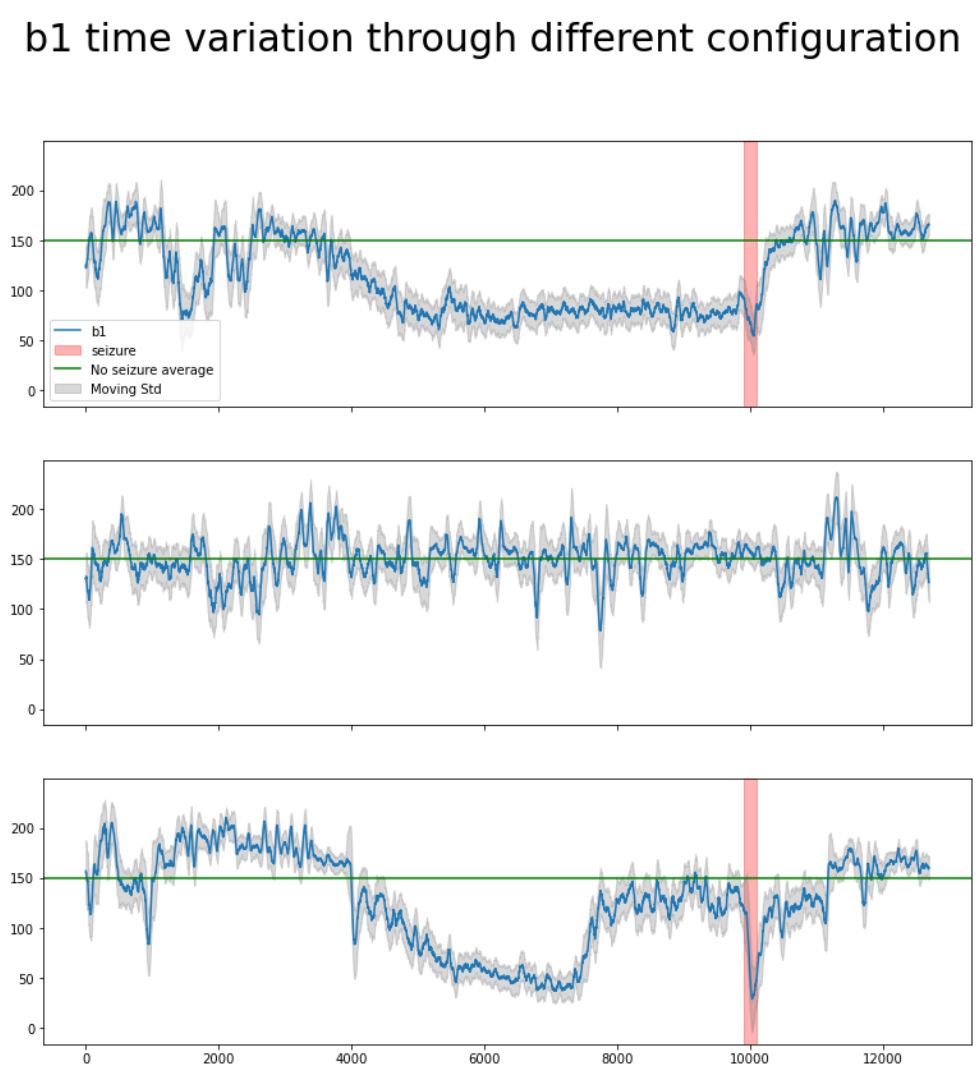



FIGURE 1 – $b_1(t)$ on 3 trajectories

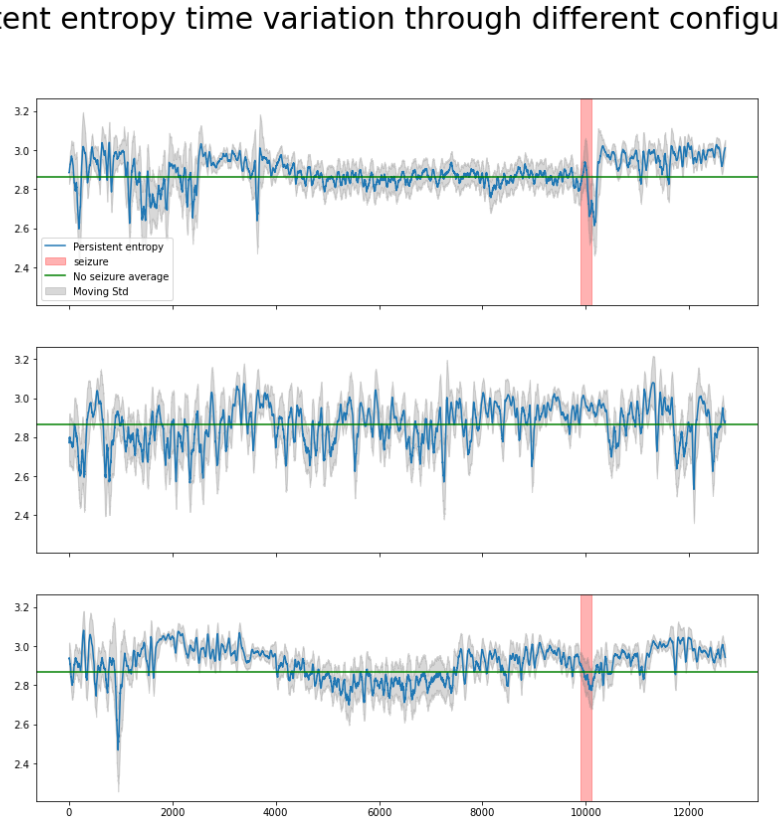



FIGURE 2 – $PE(t)$ on 3 trajectories

**General behavior (second trajectory) :** We see a volatility around the average value along time, with small perturbations that we cannot explain. The volatility goes from 100 to 200 for $b_1$ and from 2.6 to 3 for $PE$.

**Critical behavior (red area in first and third trajectory) :** The critical behavior is caracterized by an abrupt diminution to 50 in $b_1$'s trajectory (more connected structure in the simplicial complex before getting back to near the average value. Unfortunately, it cannot really be distinguished from $PE$.

**Pre-critical behavior (6000 times before critical) :** The precritical behavior is characterised by a steep decrease of $b_1$ (between 50 and 100) on the same time interval for the two seizure trajectories. An sudden increase in noticeable just before the next seizure, whilst remaining under the average value.
Heuristically, we can observe a decrease of the $PE$'s standard deviation during the precritical stage. We plan to extend this study in a future work.

## 2.3 Discussion

Many improvements can still be made to the proposed methodology, in particular (i) using Knockoff filters to guarantee that the faces are selected with confidence, (ii) using SLOPE instead of the LASSO, (iii) using higher-order sub-complexes, etc. One may also use the path

$(b_1(t), PE(t))$ and its signature as well for extracting more sensitive detection pipelines.

# Références


[1] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis : fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4 :108, 2021.

[2] Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized baker-hausdorff formula. *Annals of Mathematics*, 65(1) :163–178, 1957.

[3] Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv :1603.03788*, 2016.

[4] Harish Chintakunta, Thanos Gentimis, Rocio Gonzalez-Diaz, Maria-Jose Jimenez, and Hamid Krim. An entropy-based persistence barcode. *Pattern Recognition*, 48(2) :391–401, 2015.

[5] Stéphane Chrétien, Ben Gao, Astrid Thébault Guiochon, and Rémi Vaucher. Leveraging the power of signatures for the construction of topological complexes for the analysis of multivariate complex dynamics.

[6] Yu-Min Chung and Austin Lawson. Persistence curves : A canonical framework for summarizing persistence diagrams. *Advances in Computational Mathematics*, 48(1) :6, 2022.

[7] Peter K Friz and Nicolas B Victoir. *Multidimensional stochastic processes as rough paths : theory and applications*, volume 120. Cambridge University Press, 2010.

[8] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet : components of a new research resource for complex physiologic signals. *circulation*, 101(23) :e215–e220, 2000.

[9] Terry Lyons and Zhongmin Qian. *System control and rough paths*. Oxford University Press, 2002.

[10] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. 2006.

[11] V Navarro, M Le Van Quyen, S Clemenceau, C Adam, C Petitmengin, F Dubeau, J Gotman, J Martinerie, and M Baulac. Seizure prediction : from myth to reality. *Revue Neurologique*, 167(3) :205–215, 2010.

[12] Giovanni Petri, Paul Expert, Federico Turkheimer, Robin Carhart-Harris, David Nutt, Peter J Hellyer, and Francesco Vaccarino. Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101) :20140873, 2014.

[13] Matteo Rucco, Filippo Castiglione, Emanuela Merelli, and Marco Pettini. Characterisation of the idiotypic immune network through persistent entropy. In *Proceedings of ECCS 2014 : European Conference on Complex Systems*, pages 117–128. Springer, 2016.

[14] Hal Schenck. *Algebraic Foundations for Applied Topology and Data Analysis*, volume 1. Springer Nature, 2022.

# Modèles mixtes

# MODÈLE CONJOINT DE PROGRESSION DE MALADIE AVEC RECALIBRATION TEMPORELLE INDIVIDUELLE

Tiphaine Saulnier[1,a], Alexandra Foubert-Samier[2,abc] & Cécile Proust-Lima[3,ab]

[a] *Univ. Bordeaux, BPH Research Center, Inserm U1219 ; Bordeaux, France*
[b] *Inserm, CIC1401-EC ; Bordeaux, France*
[c] *CHU Bordeaux, Service de Neurologie des Maladies Neurodégénératives, IMNc, CRMR AMS, BPH Research Center ; Bordeaux (France)*

[1] *tiphaine.saulnier@u-bordeaux.fr*
[2] *alexandra.samier-foubert@u-bordeaux.fr*
[3] *cecile.proust-lima@u-bordeaux.fr*

**Résumé.** Établir l'histoire naturelle d'une maladie permet de mieux comprendre sa progression au cours du temps. Cependant, dans le cas de maladies complexes, son étude se confronte souvent à de multiples challenges. Lorsque la maladie est difficile à diagnostiquer, une incertitude persiste quant au moment du début de maladie. Dans les cohortes, les patients sont potentiellement recrutés à différents stades de maladie, rendant le temps d'étude moins pertinent. La survenue d'événements cliniques, tels que le décès, perturbe voire interrompt également les suivis, induisant des données manquantes potentiellement non aléatoires. Ce travail introduit un modèle conjoint combinant un modèle de progression de maladie basé sur une recalibration temporelle individuelle pour décrire la progression de marqueurs en fonction du temps de maladie latent et un modèle de survie pour évaluer l'association avec le risque de décès pendant le suivi. Cette méthodologie est motivée par l'étude de l'atrophie multi-systémique (AMS), une maladie neurodégénérative rare. Les progressions des marqueurs sont décrites en fonction du temps de maladie à l'aide de modèles à effets mixtes non linéaires. Le temps de maladie est défini en fonction du score de sévérité de la maladie à l'inclusion et d'un décalage individuel aléatoire. Les risques d'événements en compétition sont modélisés conjointement par un modèle spécifique à la cause à risques proportionnels incluant la dynamique des marqueurs et le décalage temporel individuel. L'estimation, implémentée dans le package R LTSM, est réalisée par maximum de vraisemblance en utilisant les approximations de quasi-Monte-Carlo pour le calcul de la vraisemblance et l'algorithme de Marquardt-Levenberg pour la maximisation. Les données annuelles de 663 patients de la cohorte AMS française ont été analysées sur 10,8 ans. La progression clinique de l'AMS a été décrite par deux scores mesurant l'atteinte fonctionnelle et motrice. Une fois recalibrées temporellement, leurs progressions s'étendaient sur 12 ans. Comparativement aux patients non dépendants à l'inclusion, les écarts temporels moyens entre les patients modérément dépendants et les patients complètement dépendants à l'inclusion étaient de 2,56 (IC à 95 % = 2.36, 2.76) et 5,84 (IC à 95 % = 4.92, 6.77) années, respectivement. Le risque de décès dépendait fortement de la dynamique des marqueurs et du décalage individuel (avec un risque plus élevé pour les patients plus avancés). Cette approche par temps de maladie latent a le potentiel de décrire la progression de maladies complexes pour lesquelles il est difficile d'établir le temps exact de début de maladie, tout en tenant compte de l'hétérogénéité des profils des patients et des sorties d'étude potentiellement informatives.

**Mots-clés.** Biostatistique, données longitudinales, recalibration temporelle individuelle, progression de maladie, sortie d'étude informative, modèle conjoint, risques compétitifs

**Abstract.** Establishing the natural history of a disease permits to better understand its progression over time. However, in complex diseases, its study often faces multiple challenges. When the disease is difficult to diagnose, uncertainty remains around the time of disease onset. Patients are potentially recruited in cohorts at different disease stages, making time in study no longer meaningful. Occurrence of clinical events, such as death, also interrupts follow-ups, inducing missing data potentially not at random. The present work introduces a joint model combining a disease progression model based on an individual temporal recalibration to describe markers progression according to the latent disease time and a survival model to assess the association with death. The methodology is motivated by the study of Multiple system atrophy (MSA), a rare neurodegenerative disease. The markers' progressions are described according to disease time using nonlinear mixed-effect models. Disease time is defined according to disease severity score at inclusion and an individual random shift. The risks of competing events are jointly modeled with cause-specific proportional hazard models including the markers' dynamics and the individual temporal shift. Estimation, implemented in the R-package LTSM, is carried out by Maximum Likelihood using Quasi-Monte-Carlo approximations for likelihood computation and Marquardt-Levenberg optimizer for maximization. Annual data of 663 patients from the French MSA cohort were analyzed over 10.8 years. MSA clinical progression was described by two scores assessing functional and motor impairments. Once time recalibrated, their progressions spanned over 12 years. Compared to non-dependent patients at inclusion, mean time gaps between moderately-dependent and helpless patients at inclusion were 2.56 (95%CI=2.36,2.76) and 5.84 (95%CI=4.92,6.77) years, respectively. Risk of death highly depended on markers' dynamics and individual shift (with higher risk for more advanced patients). This latent disease time approach has potential to describe the progression of complex disease for which it is difficult to establish the exact time of disease onset, while accounting for heterogeneity of patients' profiles and informative dropout.

**Keywords.** Biostatistic, longitudinal data, individual time recalibration, disease progression, informative dropout, joint model, competing risks

# 1 Introduction

Établir l'histoire naturelle d'une maladie permet de mieux comprendre sa progression au cours du temps et d'estimer correctement le pronostic des patients. Cependant, dans le cas de maladies complexes telles que les maladies neurodégénératives, son étude se confronte souvent à de multiples challenges.[1] Lorsque la maladie est difficile à diagnostiquer, une incertitude persiste quant au moment exact du début de la maladie.[2] Dans les cohortes, les patients sont potentiellement recrutés à différents stades de la maladie ou présentent des profils de progression hétérogènes, rendant le temps d'étude moins pertinent. De plus, la survenue d'événements cliniques, tels que le décès, peut perturber voire interrompre les suivis, induisant des données manquantes potentiellement non aléatoires.[3]

Pour relever ces challenges, des modèles de progression de la maladie (*disease progression model* - DPM) basés sur une recalibration temporelle ont été développés permettant de redéfinir le temps de la maladie à partir des trajectoires des patients et décrire la progression de plusieurs marqueurs de la maladie le long du continuum de la maladie.[4]

Ce travail étend cette méthodologie à la prise en compte d'un événement clinique associé (e.g., le décès) en combinant un modèle de progression de maladie basé sur une recalibration

temporelle individuelle pour décrire la progression des marqueurs en fonction du temps de maladie latent et un modèle de survie pour évaluer l'association avec le risque de décès pendant le suivi. Cette méthodologie est motivée par l'étude de l'atrophie multi-systémique (AMS), une maladie neurodégénérative rare.[5]

# 2 Méthodologie

Le principe du modèle conjoint développé dans ce travail et appliqué à l'AMS est résumé en **Figure 1**.

Le modèle décrit les trajectoires de marqueurs de progression durant le cours de la maladie. Cependant, au lieu d'utiliser le temps classique observé depuis l'inclusion, $t_{ijk}$ pour le marqueur *k* (= 1, …, K), la visite *j* (= 1, …, $n_i$) et le patient *i* (= 1, …, N), une recalibration temporelle en « temps latent de maladie » $t_{ijk} + s_i$ est réalisée. Le recalage temporel $s_i = X_{si}^{\top}\beta_s + z_{si}$, qui repositionne les courbes de progression des patients sur le continuum de la maladie, combine un recalage au niveau du groupe (c.-à-d. en fonction du stade de maladie du patient à l'entrée dans l'étude) avec $X_s^{\top}$ un vecteur de covariables et $\beta_s$ le vecteur des paramètres associés, ainsi qu'un recalage au niveau individuel par la déviation du patient par rapport au recalage du groupe avec $z_s \sim N(0, \sigma_s^2)$ représentant le recalage individuel résiduel parmi les patients d'un même stade de maladie à l'inclusion.

La progression des marqueurs au cours du temps de maladie est modélisée par un modèle multivarié à effets mixtes. La valeur du *k*-ième marqueur $Y_{ijk}$ pour le patient *i* lors de la visite *j* est décrite par :

$$H_k(Y_{ijk}; \eta_k) = F_{ik}(t_{ijk} + s_i; \theta_{ik}) + \epsilon_{ijk} \quad (1)$$

Avec $\epsilon_{ijk} \sim N(0, \sigma_{\epsilon_k}^2)$ l'erreur de mesure. La fonction $H_k(.;\eta_k)$ est une fonction de lien paramétrique utile pour normaliser quand nécessaire des marqueurs continus non-gaussiens (par exemple, via I-splines quadratiques).[1,3] La fonction paramétrique $F_{ik}(t;\theta_{ik})$ capture la forme de la trajectoire du marqueur au cours du temps. Dans la littérature sur les modèles de progression de maladie, la fonction logistique généralisée $F_{ik}(t;\theta_{ik}) = L_k + \frac{U_k - L_k}{(1+exp(-\lambda_k \cdot t))^{\nu_{ik}}} + u_{0ik}$ est souvent utilisée pour définir une progression de forme sigmoïde (avec $L_k$ et $U_k$ les asymptotes basse et haute, $\lambda_k$ le taux de progression, $\nu_{ik} \sim N(\mu_{\nu_k}, \sigma_{\nu_k}^2)$ le paramètre individuel définissant la position sur le continuum de la maladie, et $u_{0ik} \sim N(0, \sigma_{u_k}^2)$ l'intercept aléatoire). Cependant, cette spécification du modèle repose sur de fortes hypothèses sur la forme de la trajectoire. Alternativement, la fonction $F_{ik}(t;\theta_{ik})$ peut être définie de façon plus souple et guidée par les données en utilisant une base de fonctions temporelles comme des splines: $F_{ik}(t;\theta_{ik}) = X_{ik}^{\top}(t)\beta_k + Z_{ik}^{\top}(t)u_{ik}$ avec $\beta_k$ le vecteur des paramètres des effets fixes, et $u_{ik} \sim N(0, B_k)$ le vecteur des effets aléatoires individuels normalement distribués. Par contraintes d'identifiabilité, nous supposons que les $(u_{ik})_k$ sont indépendants entre les marqueurs, ainsi le recalage temporel $s_i$ représente la seule source de corrélation inter-marqueurs.[2,6]

Le risque de chaque cause *p* (= 1, …, P) d'évènement associé est décrit en fonction du recalage temporel et de la dynamique des marqueurs par un modèle à risques proportionnels spécifique à chaque cause :

$$\lambda_{ip}(t) = \lambda_{0p}(t) \exp\left( X_{ip}^{\top} \beta_p + s_i \alpha_{sp} + \sum_{k=1}^{K} g_{ikp}^{\top}(u_{ik}, t)\, \alpha_{kp} \right) \quad (2)$$

Où la fonction de risque instantanée peut dépendre de covariables $X_{ip}$, du recalage temporel individuel $s_i$, et de la dynamique des marqueurs par K fonctions $g_{ik}(u_{ik}, t)$ (e.g., les effets aléatoires $u_{ik}$, le niveau courant du marqueur $k$).

L'estimation est réalisée par la méthode du maximum de vraisemblance via l'algorithme d'optimisation de Marquardt-Levenberg.[8] L'intégrale sur les effets aléatoires dans la vraisemblance jointe est approchée par la méthode de quasi-Monte-Carlo.[7]

La procédure d'estimation est implémentée dans le package R LTSM (https://github.com/TiphaineSAULNIER/LTSM).

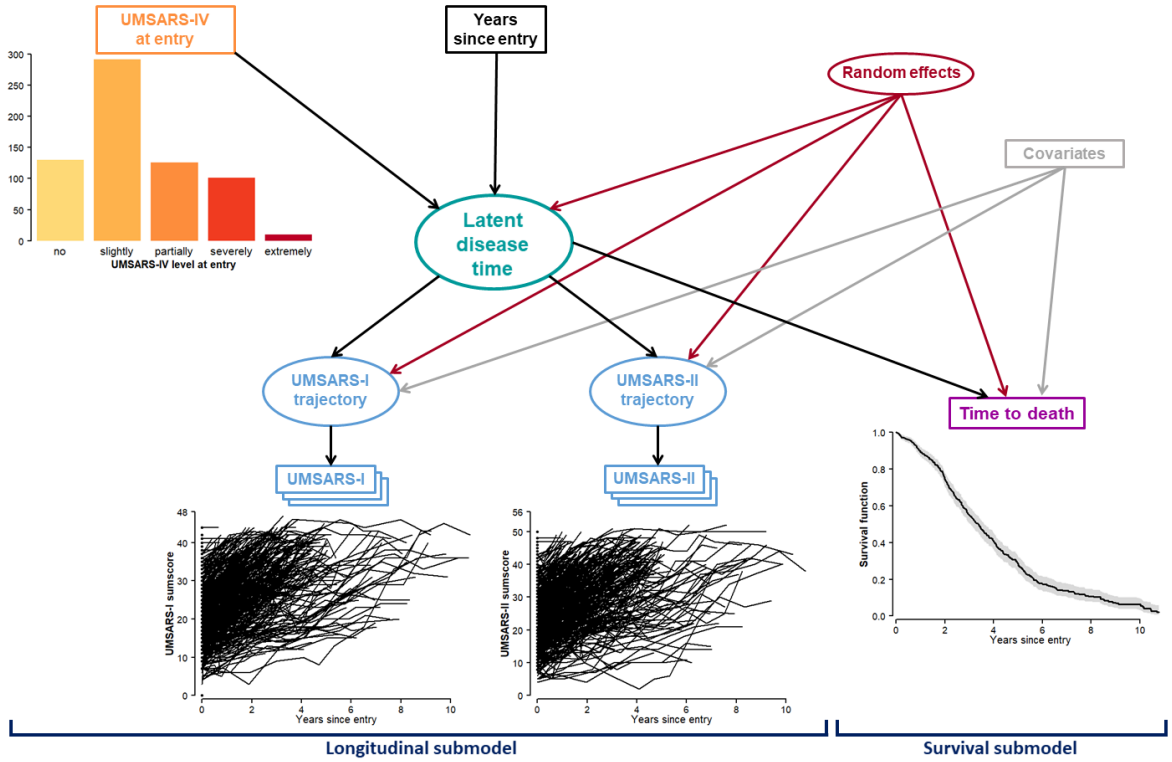



**Figure 1** : Schéma de la structure du modèle conjoint avec recalage temporel, ici sur l'application à l'AMS. La progression de l'AMS est décrite par les marqueurs UMSARS-I et UMSARS-II observés de façon répétée et modélisés par des modèles à effets mixtes. Le temps latent de maladie est défini par les années depuis l'entrée dans la cohorte recalées selon le stade de handicap (UMSARS-IV) des patients à l'entrée ainsi qu'un intercept aléatoire. L'association avec le risque de décès est évaluée par les effets aléatoires et le recalage temporel individuels. Sur ce schéma, les éléments dans les rectangles sont observés, ceux dans les ovales sont latents.

# 3 Application

Cette approche de modélisation a été utilisée pour décrire la progression de l'AMS, une alpha-synucléinopathie rare et fatale. Nous avons exploité les données de la cohorte française AMS, comprenant 663 patients suivis chaque année pendant un maximum de 10.8 ans. Les mesures répétées de deux sous-scores de l'échelle Unified MSA Rating Scale (UMSARS)[9] ont été analysées : le score UMSARS-I évaluant l'atteinte dans le domaine fonctionnel (score de 0 à 48) et le score UMSARS-II évaluant l'atteinte dans le domaine moteur (score de 0 à 56) (**Figures 2.A**). Le décalage temporel a été modélisé selon le stade de maladie à l'inclusion mesuré par le degré UMSARS-IV (**Figure 1**), une mesure globale du handicap en 5 degrés (de 1 = complètement indépendant à 5 = totalement dépendant). Les trajectoires des marqueurs ont été décrites par le modèle de l'équation (1) avec des fonctions de lien H(.) approximées par

des I-splines quadratiques et des formes linéaires de progression au cours du temps de maladie. Le risque de décès a été modélisé par l'équation (2) en fonction du décalage temporel individuel et des effets aléatoires spécifiques aux marqueurs (**Figure 1**).

La différence moyenne de temps de maladie à l'entrée dans l'étude a été estimée à 1.2, 2.6, 3.7 et 5.8 ans pour les patients avec un degré de handicap de 2, 3, 4 et 5 respectivement, par rapport aux patients sans handicap à l'inclusion (degré 1), avec de petites différences inter-patients (sd=0.6 de l'effet individuel) (**Figures 2.B**, **Figure 3.A**).

Après recalibrage temporel, les trajectoires de l'UMSARS-I et UMSARS-II (**Figures 2.C**) ont présenté une forme de progression plus homogène le long du continuum de la maladie. Pour l'UMSARS-I, la progression moyenne d'un score de 10 à un score avancé de 40 s'étend sur 9.3 ans. Pour l'UMSARS-II, la progression moyenne de 15 à 45 s'étend sur 11.6 ans.

L'âge prédit au temps de maladie 0 (correspondant à l'âge moyen d'entrée des patients qui étaient complètement indépendants à l'entrée de l'étude) a montré une très forte corrélation avec l'âge rapporté au début des symptômes ($\rho$=0,96) (**Figure 3.B**), avec le temps de maladie 0 se produisant environ 2.8 ans après la survenue rapportée des premiers symptômes.

Enfin, comme attendu, le risque de décès est fortement associé au décalage temporel individuel (risque plus élevé pour les patients plus avancés à l'entrée dans l'étude) et à la dynamique des marqueurs (risque plus élevé pour les patients avec une dégradation plus agressive que la moyenne) (**Figure 3.C**).

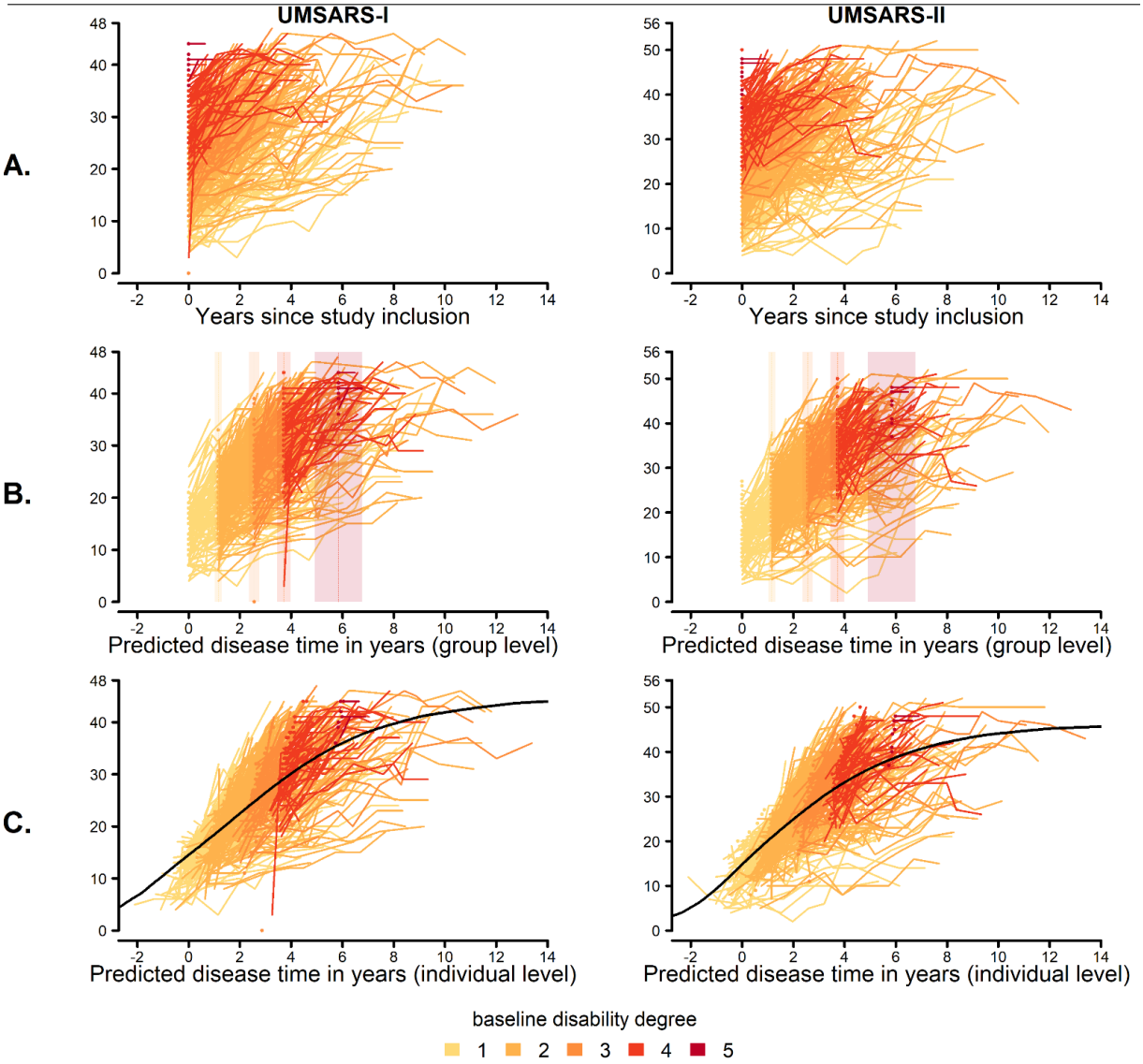



**Figure 2** : Progression de l'UMSARS-I et UMSARS-II dans la cohorte AMS française. Sont représentées ici les trajectoires individuelles des scores selon le temps observé depuis l'entrée dans la cohorte (**A.**), selon le temps de maladie prédit après recalage en fonction du stade de handicap à l'entrée (**B.**), et selon le temps individuel de maladie prédit (**C.**). Sur la ligne **B.**, le temps 0 correspond au temps d'entrée pour les patients inclus dans la cohorte sans handicap (degré 1). Sur la ligne **C.**, le temps 0 correspond au temps moyen d'entrée pour les patients inclus dans la cohorte sans handicap (degré 1).

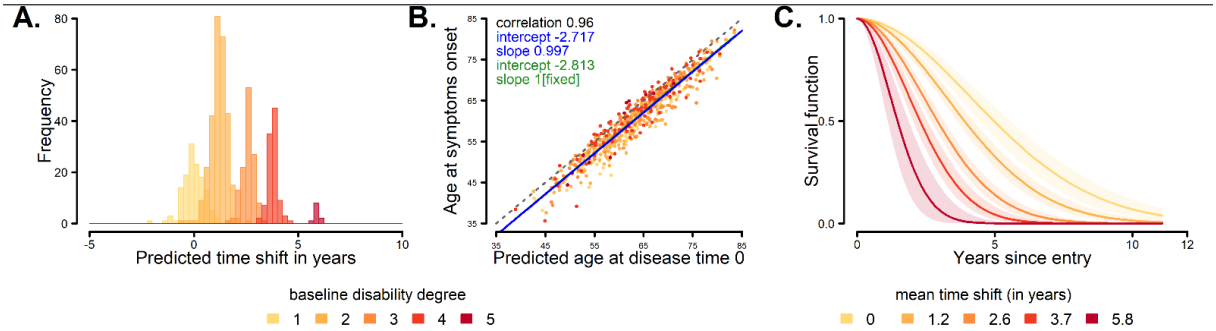



**Figure 3** : Graphes de distribution des recalages temporels individuels, de l'âge et du risque de décès dans la cohorte AMS française. Sont représentées ici la distribution du recalage temporel individuel selon le degré de handicap à l'entrée (**A.**), la correspondance entre l'âge observé lors de la survenue des premiers symptômes et l'âge prédit au temps de maladie 0 (**B.**), et les fonctions de survie selon le temps moyen de recalage de chaque stade (**C.**). Sur le graphe **B.**, deux régressions linéaires entre l'âge rapporté par le patient aux premiers symptômes versus l'âge prédit au temps de maladie 0 par le modèle ont été réalisées, une (en bleu) avec l'intercept et la pente à estimer (une pente à 1 montre une parfaite correspondance entre les âges), et une autre (en vert, confondue avec la bleue) où la pente est fixée à 1 (l'intercept quantifie le délai entre les deux âges, par exemple intercept=2 signifie que le temps de maladie survient 2 années après les premiers symptômes).

# 4 Conclusion

L'hétérogénéité temporelle dans les cohortes observationnelles constitue une barrière majeure à la compréhension de la progression des maladies. En combinant un modèle conjoint pour les sorties d'étude informatives induites par le décès et un modèle à temps de maladie latent, cette approche, et le logiciel LTSM sous R associé, offrent une solution avec un potentiel dans de nombreuses maladies complexes.

# Bibliographie


1 Commenges D, Jacqmin-Gadda H. *Dynamical Biostatistical Models*. New York: Chapman and Hall/CRC 2015. https://doi.org/10.1201/b19109

2 Saulnier T, Fabbri M, Pavy-Le Traon A, *et al.* Disease Progression in Multiple System Atrophy: The Value of Clinical Cohorts with Long Follow-Up. *Movement Disorders*. 2023;38:1567–9.

3 Saulnier T, Philipps V, Meissner WG, *et al.* Joint models for the longitudinal analysis of measurement scales in the presence of informative dropout. *Methods*. 2022;203:142–51.

4 Li D, Iddi S, Thompson WK, *et al.* Bayesian latent time joint mixed effect models for multicohort longitudinal data. *Stat Methods Med Res*. 2019;28:835–45.

5 Foubert-Samier A, Pavy-Le Traon A, Guillet F, *et al.* Disease progression and prognostic factors in multiple system atrophy: A prospective cohort study. *Neurobiol Dis*. 2020;139:104813.

6 Kühnel L, Raket LL, Åström DO, *et al.* Disease Progression in Multiple System Atrophy—Novel Modeling Framework and Predictive Factors. *Movement Disorders*. 2022;37:1719–27.

7 Pan J, Thompson R. Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis*. 2007;51:5765–75.

8 Philipps V, Hejblum BP, Prague M, *et al.* Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg. *The R Journal*. 2021;13:273.

9 Wenning GK, Tison F, Seppi K, *et al.* Development and validation of the Unified Multiple System Atrophy Rating Scale (UMSARS). *Mov Disord*. 2004;19:1391–402.

# Estimation et sélection de variables dans un modèle joint de survie et de données longitudinales avec des effets aléatoires.

Antoine Caillebotte [1,2] & Estelle Kuhn [2] & Sarah Lemler [3]

[1] *Université Paris-Saclay, INRAE, UMR GQE-Moulon, France, caillebotte.antoine@inrae.fr,*
[2] *Université Paris-Saclay, INRAE, UR MaIAGE, France, estelle.kuhn@inrae.fr,*
[3] *Université Paris-Saclay, CentraleSupélec, Laboratoire MICS, France, sarah.lemler@centralesupelec.fr*

**Résumé.** Ce travail se concentre sur l'étude jointe d'un modèle de survie et d'un modèle à effets mixtes pour expliquer le temps de survie à partir de données longitudinales et de covariables de grande dimension. Les données longitudinales sont modélisées à l'aide d'un modèle non linéaire à effets mixtes, dans lequel la fonction de régression sert de fonction de lien incorporée dans un modèle de Cox en tant que covariable. De cette manière, les données longitudinales sont liées à la durée de survie à un moment donné. De plus, le modèle de Cox prend en compte l'inclusion de covariables de grande dimension. Les principaux objectifs de cette recherche sont doubles : premièrement, identifier les covariables pertinentes qui contribuent à expliquer le temps de survie, et deuxièmement, estimer tous les paramètres inconnus du modèle joint. Pour ce faire, nous considérons la maximisation de la vraisemblance pénalisée par le LASSO. Pour résoudre le problème d'optimisation, nous introduisons un gradient stochastique préconditionné adapté aux variables latentes du modèle non linéaire à effets mixtes, associé à un opérateur proximal qui permet de gérer la non-différentiabilité de la pénalité. Nous proposons une large étude de simulations afin de montrer les performances de la procédure proposée, à la fois en termes de sélection de variables et d'estimation des paramètres du modèle considéré.

**Mots-clés.** Statistique appliquée, grande dimension et réduction de dimension, Données de survie, données censurées, biostatistique, statistique computationnelle.

**Abstract.** This paper considers a joint survival and mixed-effects model to explain the survival time from longitudinal data and high-dimensional covariates. The longitudinal data is modeled using a nonlinear mixed effects model, where the regression function serves as a link function incorporated into a Cox model as a covariate. In that way, the longitudinal data is related to the survival time at a given time. Additionally, the Cox model takes into account the inclusion of high-dimensional covariates. The main objectives of this research are two-fold: first, to identify the relevant covariates that contribute to explaining survival time, and second, to estimate all unknown parameters of the joint model. For that purpose, we consider the maximization of a LASSO penalized likelihood. To tackle the optimization problem, we implement a pre-conditioned stochastic gradient to handle the latent variables of the nonlinear mixed-effects model associated with a proximal operator to manage the non-differentiability of the penalty. We provide an extensive simulation study showcasing the performance of the proposed variable selection and the parameter estimation method.

**Keywords.** Applied statistics, high dimension and dimension reduction, survival data, censored data , biostatistics, computational statistics.

# 1 Introduction

Une problématique très actuelle dans de nombreux domaines consiste à mieux comprendre les interactions entre des phénomènes dynamiques dépendants. On peut considérer, en médecine, la dynamique des tumeurs d'un patient en oncologie et les effets des traitements anticancéreux administrés au patient. Les phénomènes considérés sont souvent complexes, tant d'un point de vue de leurs modes d'interaction que de leur dynamique temporelle et spatiale. De plus, ces phénomènes sont souvent observés dans des populations d'individus hétérogènes ou structurés.

La modélisation mathématique s'est avérée être un outil puissant pour comprendre les interactions entre plusieurs phénomènes dynamiques. Elle permet également de prendre en compte la variabilité présente dans la population observée d'individus. La modélisation jointe de plusieurs phénomènes a en particulier démontré son efficacité dans plusieurs domaines, notamment la médecine, la pharmacologie et la biologie [Kerioui *et al.*, 2022]. Un cas particulier de modèles joints concerne la modélisation simultanée de données longitudinales et de données de survie observées sur le même individu. Dans ce type de modèle joint, les données longitudinales sont souvent modélisées par un modèle à effets mixtes [Davidian et Giltinan, 1995], et les données de survie par un modèle de Cox [Cox, 1972]. Ce dernier permet de modéliser le risque instantané de la variable de survie en fonction de covariables. La modélisation des données longitudinales intervient en tant que covariable dans le modèle de Cox via une fonction de lien. L'objectif est alors d'estimer les paramètres du modèle à partir des observations et de sélectionner les covariables pertinentes [Rizopoulos, 2012]. En raison de la présence de variables latentes dans le modèle à effets mixtes, l'inférence par maximum de vraisemblance est délicate à réaliser et nécessite d'être adaptée, par exemple via les algorithmes de type Expectation Maximization (EM) [Rizopoulos, 2012]. Les algorithmes de type EM, tels que le Stochastic Approximation Expectation Maximization (SAEM), sont les approches les plus classiques pour inférer les paramètres en présence de variables latentes. Ils sont particulièrement faciles à mettre en œuvre dans le contexte d'une famille exponentielle courbe basée sur des statistiques exhaustives du modèle. De plus, des résultats de convergence théoriques de l'algorithme ont été établis dans ce contexte. Cependant, lorsque le modèle n'appartient pas à la famille exponentielle, ce qui est le cas dans notre contexte, l'algorithme est difficile à mettre en œuvre et les garanties théoriques ne sont pas établies.

Les méthodes basées sur le gradient, souvent omises, mais pourtant adaptées à l'estimation des paramètres dans les modèles latents, ne nécessitent pas d'être dans un modèle de la famille exponentielle. Ainsi, [Baey *et al.*, 2023] a suggéré d'utiliser un algorithme de gradient stochastique préconditionné pour traiter l'estimation des paramètres en présence de variables latentes. À noter que des méthodes numériques bayésiennes ont également été proposées en parallèle [Rizopoulos, 2012], [Kerioui *et al.*, 2022].

Par ailleurs, dans de nombreuses applications, les moyens technologiques actuels permettent de collecter des covariables explicatives de grande dimension. Celles-ci peuvent être, par exemple, des marqueurs génétiques ou des données omiques. Au-delà de la richesse d'informations fournies par ces covariables, elles génèrent des difficultés dans l'analyse statistique des modèles, car il est nécessaire d'adapter les approches statistiques et numériques à leur grande dimension. Une approche possible est de considérer un estimateur pénalisé, tel que le LASSO [He *et al.*, 2015], et des méthodes numériques adaptées, telles que le gradient proximal stochastique [Achab, 2017, Fort *et al.*, 2017].

Nous considérons dans cette contribution un modèle joint qui combine, par le biais d'une fonction de lien, un modèle non linéaire à effets mixtes pour les données longitudinales et un modèle de Cox pour les temps de survie, incluant des covariables de grande dimension. Notre travail vise à sélectionner les variables pertinentes parmi les covariables de grande dimension dans la partie survie du modèle joint sur la base de l'ensemble des données et ensuite d'estimer les paramètres inconnus du modèle. À cette fin, nous utilisons un gradient proximal stochastique préconditionné pour traiter les variables latentes dans le modèle joint et la pénalisation LASSO. L'algorithme proposé est facile à mettre en œuvre dans des modèles joints généraux sans supposer que la densité du modèle appartient à la famille exponentielle courbe.

Le modèle joint est détaillé dans la section 2. Dans la section 3, nous présentons la méthode d'inférence proposée basée sur un estimateur pénalisé par LASSO et une procédure numérique basée sur un algorithme de gradient proximal stochastique. Enfin, nous illustrons la méthodologie à la section 4 par une étude de simulation.

# 2 Modèle joint pour données de survie et longitudinales

Nous considérons $N$ individus et étudions, pour chaque individu $i$, le temps de survie $\mathbf{T}_i$, correspondant à la durée jusqu'à la survenue d'un événement d'intérêt, et des données longitudinales, plus précisément des observations répétées $J$ fois notées $\mathbf{Y}_{i,j}$ avec $i \in \{1, \ldots, N\}$ et $j \in \{1, \ldots, J\}$. Soit $\mathcal{D}_i = (\mathbf{Y}_i, \mathbf{T}_i, \delta_i)$ les variables observées.

## 2.1 Modèle de Survie

Le temps de survie $\mathbf{T}_i$ de l'individu $i$ est le temps entre un instant initial et la survenue d'un événement d'intérêt et est modélisé par une variable aléatoire positive. Pour caractériser la distribution de $\mathbf{T}_i$, nous utilisons la fonction de risque définie par :

$$h_i(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq \mathbf{T}_i < t + \Delta t | \mathbf{T}_i \geq t)}{\Delta t}; \forall t \geq 0. \tag{1}$$

Le modèle de Cox [Cox, 1972] est l'un des modèles les plus classiques en analyse de survie. Il relie la fonction de risque du temps de survie $\mathbf{T}_i$ aux covariables $U_i \in \mathbb{R}^p$, avec $p$ le nombre

de covariables. Le modèle de Cox pour l'individu $i$ est écrit comme suit :

$$h(t|U_i) = h_{\theta_b}(t)\exp(\beta^T U_i), \tag{2}$$

avec $\beta \in \mathbb{R}^p$ un paramètre de régression et $h_b$ la fonction de risque de base qui caractérise un comportement commun dans la population observée. Nous considérerons un risque de base paramétrique où $\theta_b$ est le vecteur de paramètres. Par conséquent, les paramètres inconnus du modèle de Cox sont $\beta$ et $\theta_b$.

En plus des covariables, nous souhaitons expliquer une partie de la variabilité du risque en utilisant la dynamique des données longitudinales, qui sera modélisée à l'aide d'un modèle à effets mixtes non linéaire. Nous présentons d'abord le modèle à effets mixtes avant d'expliquer l'intégration de cette nouvelle composante dans le modèle de Cox.

## 2.2 Modèle Non Linéaire à Effets Mixtes

Les données longitudinales sont observées $J$ fois pour chaque individu $i \in \{1, \dots, N\}$. Notons par $\mathbf{Y}_{i,j}$ la $j$-ème observation de l'individu $i$ pour $j \in \{1, \dots, J\}$ et $i \in \{1, \dots, N\}$. Nous modélisons cette observation longitudinale à l'aide d'une fonction non linéaire $m$ qui dépend des paramètres individuels représentés par la variable latente $Z_i$ comme suit :

$$\begin{cases} \mathbf{Y}_{i,j} & = m(t_j; Z_i) + \varepsilon_{i,j}, \\ Z_i & \underset{i.i.d.}{\sim} \mathcal{N}(\mu, \Gamma) \ ; \ \varepsilon_{i,j} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \end{cases} \quad \forall 1 \leq i \leq N, 1 \leq j \leq J \tag{3}$$

où, $t_j$ est le temps de la $j$-ème observation, et $\varepsilon_{i,j}$ est un bruit additif supposé centré gaussien avec une variance inconnue $\sigma^2$. La variable latente $Z_i$ décrit la variabilité interindividuelle de la population. On suppose que $Z_i$ suit une distribution gaussienne avec une espérance inconnue $\mu$ et une variance $\Gamma$ inconnue. Les paramètres inconnus du modèle à effets mixtes non linéaire sont donc $\mu, \Gamma$, et $\sigma^2$.

Nous introduisons ensuite la fonction de lien, qui combine les deux modèles précédents en modélisant l'influence de la dynamique de l'observation longitudinale sur la fonction de risque.

## 2.3 Modèle Joint de Survie et Longitudinal à Effets Mixtes

Nous supposons que le risque du temps de survie est lié à la dynamique des données longitudinales à travers la fonction de lien au sein du modèle joint définie par :

$$\begin{cases} h(t|\mathcal{M}(t, Z_i), U_i) & = h_{\theta_b}(t)\exp(\beta^T U_i + \alpha m(t, Z_i)) \\ Y_{i,j} & = m(t_j; Z_i) + \varepsilon_{i,j} \\ Z_i \underset{i.i.d.}{\sim} \mathcal{N}(\mu, \Gamma) & \ ; \ \varepsilon_{i,j} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \end{cases} \quad \forall 1 \leq i \leq N, 1 \leq j \leq J \tag{4}$$

où $\mathcal{M}(t; Zi) = \{m(s; Zi) | \forall s, 0 \leq s < t\}$ décrit les valeurs passées de la dynamique longitudinale jusqu'au temps $t$. Le paramètre $\alpha$ représente l'influence de la dynamique longitudinale sur les données de survie.

Les paramètres inconnus du modèle joint comprennent les paramètres du modèle de Cox et ceux du modèle à effets mixtes non linéaires, ainsi que le paramètre de fonction de lien du modèle joint. Nous notons $\theta = (\theta_b, \beta, \mu, \Gamma, \sigma^2, \alpha) \in \Theta$ le vecteur des paramètres inconnus avec $\Theta \subset \mathbb{R}^d$ l'espace des paramètres.

# 3 Inférence des paramètres

Dans cette section, nous proposons une méthode d'estimation des paramètres du modèle présenté ci-dessus.

## 3.1 Définition de la Vraisemblance Marginale

Nous considérons l'estimateur du maximum de vraisemblance pour estimer les paramètres du modèle joint. Dans le contexte des modèles à variables latentes, on considère la vraisemblance marginale, définie par :

$$\mathcal{L}_{marg}(\theta; \mathcal{D}) = \prod_{i=1}^{n} \int p_\theta(\mathcal{D}_i, Z_i) dZ_i \tag{5}$$

où $p_\theta(\mathcal{D}, Z)$ est la densité du couple $(\mathcal{D}, Z)$.

En raison de l'intégrale, il est difficile de calculer directement le maximum de la vraisemblance marginale, qui n'a pas de forme analytique dans ce modèle de variable latente. Par conséquent, nous utilisons des méthodes numériques pour résoudre ce problème de maximisation.

## 3.2 Définition de l'estimateur pénalisé pour la sélection de variables

Nous introduisons une pénalité et considérons un estimateur du maximum de vraisemblance pénalisé pour traiter la grande dimension des covariables. Notre objectif est de sélectionner les variables pertinentes parmi les covariables du modèle de survie. Nous utilisons la procédure LASSO (Least Absolute Shrinkage and Selection Operator) qui a été initialement développée pour les modèles de régression linéaire et le modèle de Cox [Tibshirani, 1997].

Nous choisissons de ne pénaliser que le vecteur de paramètres $\beta$ : $\text{pen}_{LASSO}(\theta) = \|\beta\|_1 = \sum_{k=1}^{p} |\beta_k|$. Nous définissons l'estimateur du maximum de vraisemblance pénalisé par :

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \left\{ \log L_{marg}(\theta; \mathcal{D}) - \lambda \ \text{pen}_{LASSO}(\theta) \right\}, \tag{6}$$

où $\Theta$ représente l'espace des paramètres et où $\lambda$ est un paramètre positif appelé paramètre de régularisation. Plus la valeur de $\lambda$ est grande, plus $\beta$ sera contraint d'avoir des composantes nulles. Inversement, plus la valeur de $\lambda$ est petite, plus les composantes de $\beta$ seront libres.

Les méthodes classiques utilisées pour estimer les paramètres sont des algorithmes de type Expectation Maximization. Ces procédures sont bien adaptées aux modèles appartenant à la famille exponentielle, ce qui n'est pas le cas du modèle joint considéré. Récemment, [Baey *et al.*, 2023] a présenté une descente de gradient stochastique préconditionnée pour l'estimation dans des modèles à variables latentes. De plus, en raison de la non-différentiabilité de la pénalité considérée, nous utilisons un algorithme proximal tel que présenté par [Achab, 2017] et [Fort *et al.*, 2017]. Ainsi, nous intégrons un gradient proximal dans la procédure présentée dans [Baey *et al.*, 2023]. Au final nous mettons en œuvre un algorithme de gradient proximal stochastique préconditionné pour calculer l'estimateur.

## 3.3 Algorithme d'Estimation

Nous mettons en œuvre un gradient proximal stochastique préconditionné, appelé SPG-FIM dans la suite. L'algorithme est divisé en trois étapes : une réalisation des variables latentes est échantillonnée lors d'une première étape appelée *Simulation*, qui utilise un algorithme de Metropolis-Hastings. La deuxième étape est la descente classique du gradient sur la vraisemblance complète approximée, l'étape *Forward*. Suivant la procédure présentée dans [Baey *et al.*, 2023], nous avons choisi d'utiliser un préconditionnement du gradient avec une estimation de la matrice d'information de Fisher (FIM). Cette dernière est mise à jour au cours des itérations à l'aide de l'estimation présentée par [Delattre et Kuhn, 2023]. La dernière étape, appelée *Backward*, traite le terme de pénalité. Nous appliquons l'opérateur proximal classique, défini ci-dessous.

$$\text{Prox}_{\text{pen}}(\beta) = \arg\min_{\beta' \in \mathbb{R}^p} \left( \text{pen}(\beta') + \frac{1}{2} \left\| \beta - \beta' \right\|_2^2 \right). \tag{7}$$

Avec la pénalité LASSO, l'opérateur proximal a la forme explicite suivante :

$$(\text{Prox}_{LASSO}(\beta))_i = \begin{cases} 0 & \text{if } |\beta_i| < \lambda \\ \beta_i - \lambda & \text{if } \beta_i \geq \lambda \\ \beta_i + \lambda & \text{if } \beta_i \leq -\lambda \end{cases} ; \ \forall i \in \{1, ..., p\}. \tag{8}$$

L'étape *Backward* correspond à l'application de l'opérateur proximal sur le résultat de l'étape *Forward*. L'Algorithme 1 détaille les étapes de SPG-FIM. Il convient de noter qu'il est possible de différencier les deux suites de taille de pas impliquées dans l'approximation stochastique ou la descente de gradient. Toutefois, par souci de clarté, nous les avons notées de la même manière ici.

Comme la pénalité ne dépend que de $\beta$, l'opérateur proximal sélectionne les composantes de $\beta$ qui semblent les plus explicatives des données. Il calcule une solution parcimonieuse pour $\beta$ mais applique également un rétrécissement sur les composantes non nulles, de sorte que l'estimateur LASSO est biaisé. C'est pourquoi nous détaillons dans ce qui suit une méthode permettant d'obtenir un estimateur non biaisé.

**Algorithm 1:** Gradient proximal stochastique avec préconditionnement FIM (SPG-FIM)

**Require:** Nombre d'itérations $K \geq 1$ ; séquence de pas $(\gamma_k)_{k\geq 1}$
**Initialize** Point de départ $\theta_0 \in \mathbb{R}^d$, $\Delta_0$
**for** $k = 1$ **to** $K$ **do**
  • **Étape de Simulation :**
    **Obtenir** $Z^{(k)}$ **avec une étape de Metropolis-Hastings**
  • **Calcul du gradient :** $v_k = \frac{1}{N}\sum_{i=1}^{N} \nabla \log p_{\theta_k}(\mathcal{D}_i, Z_i^{(k)})$
  • **Calcul du FIM :**
    • **Calculer l'approximation stochastique**
    $\forall i \in \{1, ..., N\}, \Delta_i^{(k)} = (1-\gamma_k)\Delta_i^{(k-1)} + \gamma_k \nabla \log p_{\theta_k}(\mathcal{D}_i, Z_i^{(k)})$
    • **Calculer le FIM :** $FIM_k = \frac{1}{N}\sum_{i=1}^{N} \Delta_i^{(k)} \left(\Delta_i^{(k)}\right)^T$
  • **Descente de gradient :**
    • **Étape Forward :** $\omega_{k+1} = \theta_k - \gamma_k FIM_k^{-1} v_k$
    • **Étape Backward :** $\theta_{k+1} = \mathrm{Prox}_{\gamma_k \mathrm{pen}}(\omega_{k+1})$
**end**
**return** $\hat{\theta} = \theta_K$

## 3.4 Procédure d'Estimation

Comme expliqué ci-dessus, l'opérateur proximal (8) a un effet rétrécissant sur l'estimateur après son application, ce qui signifie que les valeurs trouvées pour $\beta$ sont plus petites que prévu, et donc l'estimateur de $\beta$ est biaisé. Pour débiaiser l'estimateur, nous allons procéder en deux étapes. Une première étape exploratoire nous permet de sélectionner le support du vecteur $\beta$ à l'aide d'une procédure Lasso,

$$\hat{\theta}_{\mathbf{LASSO}}(\lambda) = \arg\max_{\theta\in\Theta} \left\{\log \mathcal{L}_{marg}(\theta; \mathcal{D}) - \lambda \mathrm{pen}_{LASSO}(\theta)\right\}.$$

La seconde étape consiste à maximiser la vraisemblance non pénalisée en les paramètres sélectionnés à l'étape précédente. Nous devons également sélectionner une valeur bien équilibrée pour le paramètre de régularisation, il est d'usage de déterminer la valeur de $\lambda$ par validation croisée. Mais ici, sans contexte prédictif, le paramètre de régularisation de la procédure Lasso est sélectionné à l'aide du critère BIC défini de la façon suivante:

$$BIC(\lambda) = -2\log(\mathcal{L}_{marg}(\hat{\theta}_{\mathbf{LASSO}}(\lambda); \mathcal{D})) + k\ \log(N(1+J)).$$

où k est le nombre de composantes non nulles dans $\hat{\beta}_{LASSO}$.

# 4 Étude de Simulation

Dans cette section, nous proposons d'étudier les performances de la procédure que nous venons de présenter. Nous considérons le modèle conjoint défini par 4 avec la fonction logistique classique pour le modèle non linéaire à effets mixtes, définie par :

$$m : t \mapsto \frac{Z_1}{1 + \exp\left(\frac{Z_2 - t}{\mu_3}\right)}, \tag{9}$$

Nous modélisons pour chaque individu $i$ le paramètre individuel correspondant $Z_i \in \mathbb{R}^2$ par une variable aléatoire gaussienne avec une espérance $(\mu_1, \mu_2) \in \mathbb{R}^2$ et une variance diagonale $\Gamma = diag(\gamma_1^2, \gamma_2^2)$. $\mu_3$ est considéré comme un paramètre de population inconnu. Nous considérons le risque de base comme étant une Weibull définie comme $h_{a,b}(t) = ba^{-b}t^{b-1}$, où $a$ et $b$ sont connus.

## Configuration de la simulation

Nous avons généré 50 ensembles de données selon le modèle joint présenté précédemment dans l'équation 4. Pour chaque valeur différente de $p$, nous choisissons le vecteur $\beta$ de telle sorte que les quatre premières composantes soient égales à $(-2, -1, 1, 2)$ et que le reste soit égal à zéro. Nous générons également la matrice des covariables $U$ avec $N$ lignes et $p$ colonnes, suivant une distribution uniforme $U_{i,l} \sim \mathcal{U}([-1, 1]),\ \forall i \in \{1, ..., N\}, l \in 1, ..., p$. Toutes les valeurs des paramètres sont détaillées dans la Table 1.

Table 1: Valeurs réelles des paramètres utilisées pour la simulation

| Paramètres | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\gamma_1^2$ | $\gamma_2^2$ | $\sigma^2$ | $\alpha$ |
|---|---|---|---|---|---|---|---|
| Valeur réelle | 0.3 | 90 | 7.5 | $2.5\ 10^{-3}$ | 20 | $10^{-3}$ | 11.11 |
| Paramètres | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | ... | $\beta_p$ |
| Valeur réelle | -2 | -1 | 1 | 2 | 0 | ... | 0 |

Notre objectif est de montrer la consistance numérique de l'estimateur 6, nous nous concentrons donc sur quatre scénarios où le nombre d'individus observés augmente $N \in \{50, 100, 200, 300\}$ lorsque le nombre de covariables et le nombre d'observations longitudinales sont fixes, $p = 200$, $J = 5$. Nous considérons séparément les erreurs quadratiques moyennes relatives ($rrmse$) des paramètres de grande dimension $\beta \in \mathbb{R}^p$ sélectionnés par la méthode et des autres, définis par $\nu \in \mathbb{R}^d$.

La Table 2 donne les erreurs relatives calculées sur 50 jeux de données pour les paramètres $\beta$ et $\nu$ séparément. On observe bien une diminution des erreurs relatives lorsque le nombre d'observations augmente.

Table 2: Erreurs d'estimation dans le modèle joint pour les scénarios $N \in \{50, 100, 200, 300\}$ et $p = 200, J = 5$.

| | Errors | |
|---|---|---|
| Scenarios | $rrmse(\beta)$ | $rrmse(\nu)$ |
| $N = 50$ | 0.848 | 0.122 |
| $N = 100$ | 0.575 | 0.106 |
| $N = 200$ | 0.171 | 0.042 |
| $N = 300$ | 0.124 | 0.020 |

On souhaite également à l'avenir obtenir des résultats sur la capacité de la méthode à sélectionner les variables les plus pertinentes dans un grand ensemble de covariables. Pour cela, on souhaite présenter des résultats de sensibilité et spécificité de l'algorithme que l'on propose. Pour cela, on fixera le nombre d'individus $N = 100$ et l'on étudiera les scénarios où le nombre de covariables augmente par exemple $P \in \{50, 150, 400\}$. On calculera alors la proportion de vrai positif et faux positif.

# 5 Discussions

Dans ce travail, nous avons considéré un modèle joint en couplant un modèle de survie avec un modèle non linéaire à effets mixtes via une fonction de lien. Nous avons traité la sélection de variables et l'estimation de paramètres dans ce modèle.

Une première perspective à ce travail serait d'introduire de la grande dimension également dans le modèle non linéaire à effets mixtes. Une seconde perspective intéressante consisterait à aborder la prédiction dans le contexte de la modélisation jointe. Il serait ainsi intéressant de mettre en place une méthode de prédiction du temps de survie à partir d'un début d'observation de données longitudinales.

**Remerciements.** Ce travail a été financé par le projet (Stat4Plant) ANR-20-CE45-0012.

# References

[Achab, 2017] Achab, M. (2017). Learning from sequences with point processes. Issue: 2017SACLX068.

[Baey *et al.*, 2023] Baey, Delattre, Kuhn, Leger et Lemler (2023). Efficient precondtionned stochastic gradient descent fro estimation in latent variables models. *ICML*.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220. Publisher: [Royal Statistical Society, Wiley].

[Davidian et Giltinan, 1995] Davidian, M. et Giltinan, D. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[Delattre et Kuhn, 2023] Delattre, M. et Kuhn, E. (2023). Computing an empirical Fisher information matrix estimate in latent variable models through stochastic approximation. *Computo*.

[Fort *et al.*, 2017] Fort, G., Ollier, E. et Samson, A. (2017). Stochastic proximal gradient algorithms for penalized mixed models.

[He *et al.*, 2015] He, Z., Tu, W., Wang, S., Fu, H. et Yu, Z. (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes: Variable selection in joint models. *Biometrics*, 71(1):178–187.

[Kerioui *et al.*, 2022] Kerioui, M., Bertrand, J., Bruno, R., Mercier, F., Guedj, J. et Desmée, S. (2022). Modelling the association between biomarkers and clinical outcome: An introduction to nonlinear joint models. *British Journal of Clinical Pharmacology*, 88(4):1452–1463.

[Rizopoulos, 2012] Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press.

[Tibshirani, 1997] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

# UN TEST BOOTSTRAP DE NULLITÉ DES COMPOSANTES DE LA VARIANCE DANS LES MODÈLES NON LINÉAIRES À EFFETS MIXTES

Tom Guédon [1] & Charlotte Baey [2] & Estelle Kuhn [3]

[1] *MaIAGE, INRAE, Université Paris-Saclay, Domaine de Vilvert, 78352 Jouy-en-Josas, France*
*tom.guedon@inrae.fr*

[2] *Laboratoire Paul Painlevé, Université de Lille, 59655 Villeneuve d'Asq, France*
*charlotte.baey@universite-lille.fr*

[3] *MaIAGE, INRAE, Université Paris-Saclay, Domaine de Vilvert, 78352 Jouy-en-Josas, France*
*estelle.kuhn@inrae.fr*

**Résumé.** Nous considérons dans ce travail les tests de nullité des composantes de la variance dans les modèles à effets mixtes. Nous prenons en compte la présence de paramètres de nuisance, c'est-à-dire le fait que certaines variances non testées pourraient également être égales à zéro. Deux problèmes principaux se posent dans ce contexte. Premièrement, sous l'hypothèse nulle, la valeur réelle du paramètre se situe sur la frontière de l'espace des paramètres. De plus, la localisation des paramètres de nuisance étant inconnue, la théorie asymptotique usuelle est inutilisable. Ensuite, dans le contexte spécifique des modèles à effets mixtes non linéaires, la matrice d'information de Fisher est singulière. Nous abordons ces deux points en proposant une procédure de bootstrap paramétrique basée sur le rapport de vraisemblance, qui est applicable même pour les modèles non linéaires. Nous montrons que la procédure est consistante, résolvant à la fois les problèmes de frontière et de singularité. Nous montrons, à travers une étude de simulation, que notre procédure, comparée à l'approche asymptotique, a de meilleures performances avec de petits échantillons et est plus robuste à la présence de paramètres de nuisance.

**Mots-clés.** modèles non linéaire à effets mixtes, Bootstrap, Test d'hypothèse, Matrice d'information de Fisher singulière

**Abstract.** We examine the problem of variance components testing in general mixed effects models using the likelihood ratio test. We account for the presence of nuisance parameters, i.e. the fact that some untested variances might also be equal to zero. Two main issues arise in this context leading to a non regular setting. First, under the null

hypothesis the true parameter value lies on the boundary of the parameter space. Moreover, due to the presence of nuisance parameters the exact location of these boundary points is not known, which prevents from using classical asymptotic theory of maximum likelihood estimation. Then, in the specific context of nonlinear mixed-effects models, the Fisher information matrix is singular at the true parameter value. We address these two points by proposing a shrinked parametric bootstrap procedure, which is straightforward to apply even for nonlinear models. We show that the procedure is consistent, solving both the boundary and the singularity issues. We show through a simulation study that, compared to the asymptotic approach, our procedure has a better small sample performance and is more robust to the presence of nuisance parameters.

**Keywords.** Nonlinear mixed efffects models, Bootstrap, Hypothesis testing, Singular Fisher information matrix

# 1 Introduction

Les modèles à effets mixtes constituent un puissant outil statistique pour modéliser des études longitudinales avec des mesures répétées, ou des données présentant une structure latente inconnue. Ces modèles permettent de prendre en compte deux types de variabilités. La première est celle existant entre différents individus, la deuxième est celle due aux différentes mesures effectuées sur le même individu. Ces variabilités sont modélisées par deux types d'effets : d'une part, des effets aléatoires qui varient d'un individu à l'autre, et d'autre part, des effets fixes, communs à tous les individus de la population (voir [4], [2]).

Du point de vue de la modélisation, pouvoir distinguer parmi tous les effets ceux qui peuvent être modélisés en tant qu'effets fixes permettrait de réduire le nombre de paramètres du modèle. Cela aiderait également à mieux identifier les processus à l'origine de la variabilité observée dans la population. Cette question peut se reformuler sous la forme d'un test d'hypothèses portant sur la nullité de certaines composantes de la variance des effets aléatoires.

Pour réaliser ce test, les difficultés à prendre en compte sont les suivantes : les variances égales à zéro sont sur la frontière de l'espace des paramètres, ce qui change la distribution asymptotique des statistiques considérées. Ainsi, si des variances non testées sont nulles, on ne peut pas déterminer la distribution asymptotique de la statistique de test qui dépend de la position de ces paramètres de nuisance [5]. Ensuite nous montrons que ces paramètres de nuisance induisent une dégénérescence de la matrice d'information de

Fisher. Ces problématiques rendent le test asymptotique difficile, ou même parfois impossible, à mettre en place.

Nous proposons une procédure de bootstrap paramétrique basée sur la statistique du rapport de vraisemblance pour réaliser ce test. Nous montrons qu'en choisissant correctement le paramètre utilisé pour générer les échantillons bootstrap, nous propososons une procédure de test bootstrap consistante même en présence de paramètres de nuisance. Nous illustrons ces résultats théoriques sur des données simulées. Notre contribution réside dans l'applicabilité du test aux modèles non linéaires, et dans la prise en compte de variances d'effets aléatoires non testées et égales à zéros.

# 2 Méthodologie proposée

## 2.1 Modèles à effets mixtes

Considérons $N$ individus mesurés chacun $J$ fois, où $N$ et $J$ sont des entiers non négatifs. Notons $y_{ij}$ $(i = 1, ..., N; j = 1, ..., J)$ la $j$-ième observation du $i$-ième individu. Nous considérons le modèle non linéaire à effets mixtes suivant

$$\begin{cases} y_{ij} & = g(x_{ij}, \beta, \Lambda\xi_i) + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \xi_i & \sim \mathcal{N}(0, I_p) \end{cases}, \tag{1}$$

où $(\xi_i)_{i=1,...,N}$ et $(\varepsilon_{ij})_{i=1,..,N,j=1,...,J}$ sont des variables aléatoires mutuellement indépendantes, $g$ est une fonction non linéaire connue, $x_{ij}$ rassemble toutes les covariables de la $j$-ième observation du $i$-ième individu, $\beta \in \mathbb{R}^b$ est le vecteur des effets fixes, $\Lambda \in \mathcal{L}_p^+$ est une matrice triangulaire inférieure, paramètre d'échelle de l'effet aléatoire $\xi_i$, et $\sigma^2$ est la variance positive du bruit.

**Remarque 1** *La définition du modèle* (1) *est légèrement plus générale que la terminologie habituelle des modèles à effets mixtes [4, p. 306], qui définit* $y_{ij} = g(v_{ij}, \phi_i) + \varepsilon_{ij}$*, avec* $\phi_i = A_{ij}\beta + B_{ij}b_i$ *le* $i$*-ième paramètre individuel,* $\beta$ *le vecteur des effets fixes associé à l'effet aléatoire* $b_i \sim \mathcal{N}(0, \Gamma)$*, et où* $v_{ij}$*,* $A_{ij}$ *et* $B_{ij}$ *sont des covariables connues. Le modèle* (1) *couvre cette définition en prenant* $x_{ij} = (v_{ij}, B_{ij}, A_{ij})$ *et* $b_i = \Lambda\xi_i$*.*

## 2.2 Tests des composantes de la variance

Dans cette section, nous présentons notre procédure de test.

Soit $r \in \{1, \ldots, p\}$ le nombre de variances à tester. Sans perte de généralité, nous supposons que nous testons la nullité des $r$ dernières variances dans $\Gamma = \Lambda\Lambda^T$. Par conséquent, considérons la notation de matrice en bloc suivante

$$\Lambda = \left( \begin{array}{c|c} \Lambda_1 & 0_{(p-r)\times r} \\ \hline \Lambda_{12} & \Lambda_2 \end{array} \right),$$

où $\Lambda_1 \in \mathcal{L}_{p-r}^+$, $\Lambda_2 \in \mathcal{L}_r^+$ et $\Lambda_{12} \in \mathcal{M}_{r\times(p-r)}(\mathbb{R})$.

Nous écrivons $\theta_0$ le vrai paramètre sur lequel nous considérons le test suivant :

$$H_0 : \theta_0 \in \Theta_0 \quad \text{contre} \quad H_1 : \theta_0 \in \Theta, \tag{2}$$

où

$$\begin{aligned} \Theta_0 &= \{\theta \in \mathbb{R}^q \mid \beta \in \mathbb{R}^b, \Lambda_1 \in \mathcal{L}_{p-r}^+, \Lambda_2 = 0, \Lambda_{12} = 0, \sigma^2 \in \mathbb{R}_*^+\} \\ \Theta &= \{\theta \in \mathbb{R}^q \mid \beta \in \mathbb{R}^b, \Lambda \in \mathcal{L}_p^+, \sigma^2 \in \mathbb{R}^+\} \end{aligned}$$

où $\mathcal{L}_k^+$ est l'ensemble des matrices triangulaires inférieurs avec des coefficients diagonaux positifs ou nuls.

**Remarque 2** *Nous n'imposons pas que la diagonale de $\Lambda_1$ soit strictement positive, ce qui permet à certaines variances non testées d'être égales à zéro.*

Nous considérons la statistique du test du rapport de vraisemblance définie comme :

$$\text{LRT}(y_{1:N}) = -2\left( \sup_{\theta\in\Theta} \quad l(\theta; y_{1:N}) - \sup_{\theta\in\Theta_0} \quad l(\theta; y_{1:N}) \right).$$

où $l(\theta; y_{1:N}) = \sum_{i=1}^N \log f(y_i; \theta)$

## 2.3 Procédure de test

Nous proposons une procédure de test par bootstrap paramétrique utilisant un paramètre de bootstrap $\theta_N^*$ et $B \in \mathbb{N}^*$ échantillons bootstrap pour tester (2) avec une erreur de type I $\alpha \in [0, 1]$.

**Algorithm 1** Procédure de bootstrap paramétrique

Entrée: $c_N > 0$, $B \in \mathbb{N}^*$, $0 < \alpha < 1$
Fixer $\beta_N^* = \hat{\beta}_N$, $\Lambda_N^* = \hat{\Lambda}_N$, et ${\sigma_N^*}^2 = \hat{\sigma}_N^2$
Fixer $\Lambda_{2,N}^* = \Lambda_{12,N}^* = 0$
Fixer $[\Lambda_{1,N}^*]_{mn} = [\hat{\Lambda}_{1,N}]_{mn} \mathbb{1}_{[\hat{\Lambda}_{1,N}]_{mn} > c_N}$
Pour $b = 1, ..., B$
  Pour $i = 1, ..., N$, simuler indépendamment $\varepsilon_i^{*,b} \sim \mathcal{N}(0, {\sigma_N^*}^2 I_J)$ et $\xi_i^{*,b} \sim \mathcal{N}(0, I_p)$
  Reconstruire la $i$ème réponse du $b^{ième}$ échantillon bootstrap $y_i^{*,b} = g(x_i, \beta_N^*, \Lambda_N^* \xi_i^{*,b}) + \varepsilon_i^{*,b}$
  Calculer la statistique de test bootstrap $\text{LRT}(y_{1:N}^{*,b})$
Calculer la $p$-value bootstrap : $p_{boot} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\text{LRT}(y_{1:N}^{*,b}) > \text{LRT}(y_{1:N})}$
Rejeter $H_0$ si $p_{boot} < \alpha$

# 3 Étude théorique

Cette section résume les principaux résultats théoriques de ce travail. Nous ne détaillons pas les différentes hypothèses de régularité imposées au modèle, qu'on pourra retrouver en détail dans [3]. Dans un premier temps nous formalisons ce que nous appelons les paramètres de nuisance. Nous supposons que, en plus des dernières $r$ variances testées, $m$ variances non testées sont nulles. Sans perte de généralité, nous supposons que les dernières $m+r$ variances des paramètres individuels sont nulles, de sorte que $\Lambda_0$ est de la forme

$$\Lambda_0 = \left(\begin{array}{c|c|c} \Lambda_1^{nonuis} & 0_{(p-r-m)\times m} & 0_{(p-r-m)\times r} \\ \hline \Lambda_{12}^{nuis} & \Lambda_1^{nuis} & 0_{m\times r} \\ \hline \Lambda_{12,1} & \Lambda_{12,2} & \Lambda_2 \end{array}\right) = \left(\begin{array}{c|c|c} \Lambda_1^{nonuis} & 0_{(p-r-m)\times m} & 0_{(p-r-m)\times r} \\ \hline 0_{m\times(p-r-m)} & 0_{m\times m} & 0_{m\times r} \\ \hline 0_{r\times(p-r-m)} & 0_{r\times m} & 0_{r\times r} \end{array}\right)$$

Il est important de noter qu'en applications réelles, les $m$ lignes induisant des paramètres de nuisance sont situées à des positions inconnues dans la matrice $\Lambda$, et que les $p-m-r$ variances restantes sont strictement positives, ce qui équivaut à ce que les coefficients diagonaux de $\Lambda_1^{nonuis}$ soient strictement positifs.

Nous divisons maintenant le paramètre en $\theta = (\psi, \delta, \lambda)$, où $\lambda$ représente tous les coefficients de $\Lambda_2$ et $\Lambda_{12,2}$, $\delta$ représente les coefficients dans $\Lambda_1^{nuis}$ et $\Lambda_{12}^{nuis}$ et $\psi$ regroupe tous les autres paramètres.

La proposition suivante décrit le fait que si la matrice d'information de Fisher existe, elle présentera des blocs égaux à zéro, et sera donc singulière. Cela formalise le problème

de singularité induit par le paramètre de nuisance $\delta$.

**Proposition 1** *Sous des hypothèses de régularité, pour $k = 0, 1$, pour tout $y \in \mathbb{R}^J$, $\nabla_\delta^{2k+1} \log f(y; \theta_0) = 0$ et $\nabla_\lambda^{2k+1} \log f(y; \theta_0) = 0$. En particulier,* $\text{var}\nabla_\delta l(\theta; y_{1:N}) = 0$ *et* $\text{var}\nabla_\lambda l(\theta; y_{1:N}) = 0$.

**Proposition 2** *Sous des hypothèses de régularité :*

$$\begin{aligned} \arg\max_{\theta \in \Theta} \quad & l(\theta; y_{1:N}) = \theta_0 + o_p(1) \\ \arg\max_{\theta \in \Theta_0} \quad & l(\theta; y_{1:N}) = \theta_0 + o_p(1) \end{aligned}$$

Un choix naturel pour le paramètre bootstrap $\theta_N^*$ est l'estimateur du maximum de vraisemblance. Cependant, le bootstrap échoue en présence des paramètres de nuisance $\delta$. C'est pourquoi il est nécessaire de faire attention lors du choix de $\delta_N^*$. Pour expliquer et résoudre ce problème, nous devons d'abord dériver la vitesse de convergence de l'estimateur du maximum de vraisemblance.

**Proposition 3** *Soient $\hat{\theta}_N = (\hat{\psi}_N, \hat{\delta}_N, \hat{\lambda}_N)$ et $\tilde{\theta}_N = (\tilde{\psi}_N, \tilde{\delta}_N, 0_{d_\lambda})$ respectivement les estimateurs du maximum de vraisemblance non restreint et restreint de $\theta$. Sous des hypothèses de régularité, on a $(\hat{\psi}_N, \tilde{\psi}_N) = O_p(N^{-1/2}), (\hat{\delta}_N, \tilde{\delta}_N, \hat{\lambda}_N) = O_p(N^{-1/4})$.*

Nous pouvons maintenant énoncer le principal résultat de ce travail qui garantit la consistance de la procédure bootstrap.

**Théorème 1** *Sous des hypothèses de régularité, si $\theta_N^*$ est choisi de telle sorte que $\theta_N^* \in \Theta_0$, $\theta_N^* = \theta_0 + o_p(1)$ et $N^{1/4}\delta_N^* = o_p(1)$ alors lorsque $N \to +\infty$, il est vrai en probabilité que*

$$\text{pr}^*\text{LRT}(y_{1:N}^*) \leq t \longrightarrow \text{pr}(\text{LRT}_\infty \leq t). \tag{3}$$

Une manière de choisir $\theta_N^*$ qui satisfait l'hypothèse du théorème 1 est de suivre l'idée de [1] et de seuiller le paramètre vers 0. La proposition (4), suivante donne une procédure pour choisir $\theta_N^*$ et une justification du choix fait dans l'algorithme 1.

**Proposition 4** *Soit $(c_N)_{N \in \mathbb{N}}$ une suite telle que $\lim_{N \to +\infty} c_N = 0$ et $\lim_{N \to +\infty} N^{\frac{1}{4}} c_N = +\infty$. Soit $\hat{\theta}_N = (\hat{\psi}_N, \hat{\delta}_N, \hat{\lambda}_N)$ un estimateur du maximum de vraisemblance (restreint ou non) de $\theta_0 = (\psi_0, 0_{d_\delta}, 0_{\delta_\lambda})$. Sous des hypothèses de régularité, en choisissant $\theta_N^* = (\psi_N^*, \delta_N^*, \lambda_N^*)$ tel que: $\forall k = 1, .., d_\psi$ $\psi_{N,k}^* = \hat{\psi}_{N,k} \mathbb{1}(\hat{\psi}_{N,k} > c_N)$, $\forall k = 1, .., d_\delta$ $\delta_{N,k}^* = \hat{\delta}_{N,k} \mathbb{1}(\hat{\delta}_{N,k} > c_N)$ et $\lambda_N^* = 0_{d_\lambda}$ alors, $\theta_N^*$ vérifie l'hypothèse du théorème 1.*

# 4 Étude de simulation

Pour cette étude de simulation nous considérons tout d'abord un modèle linéaire à deux effets aléatoires:

$$y_{ij} = \beta 1 + \lambda_1 \xi_{i1} + (\beta_2 + \lambda_2 \xi_{i2}) j + \varepsilon_{ij}$$

où l'on teste la nullité de $\lambda_2$:

$$H_0 : \lambda_2 = 0 \quad \text{contre} \quad H_1 : \lambda_2 > 0$$

. La Table 1 présente les niveaux estimés pour différentes tailles d'échantillon. On compare la procédure bootstrap à la procédure asymptotique.

| Level $\alpha$ | $N = 10$ | | $N = 20$ | | $N = 30$ | | $N = 40$ | | $N = 100$ | | max sd |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | boot | asym | boot | asym | boot | asym | boot | asym | boot | asym | |
| 1% | 1.14 | 0.68 | 0.98 | 0.68 | 1.20 | 0.94 | 0.74 | 0.70 | 0.86 | 0.72 | 0.15 |
| 5% | 5.20 | 3.64 | 5.22 | 3.82 | 5.74 | 4.30 | 4.86 | 3.94 | 5.26 | 4.50 | 0.33 |
| 10% | 10.72 | 7.16 | 10.80 | 7.98 | 10.30 | 8.40 | 10.80 | 8.44 | 10.34 | 8.86 | 0.44 |

Table 1: Niveaux empiriques (en pourcentage) dans le test qu'une variance est nulle dans un modèle linéaire à deux effets aléatoires. L'expérience a été répétée $K = 5000$ fois et $B = 500$ échantillons bootstrap ont été utilisés pour chaque expérience

Dans un second temps, on évalue la procédure sur un modèle non linéaire. On considère le modèle de croissance logistique défini comme suit :

$$y_{ij} = \frac{\beta_1 + \lambda_1 \xi_{i1}}{1 + \exp\left(-\frac{x_{ij} - (\beta_2 + \lambda_2 \xi_{i2})}{\beta_3 + \lambda_3 \xi_{i3}}\right)} + \varepsilon_{ij}$$

La Table 2 montre les résultats de l'expérience. On compare la procédure bootstrap à la procédure asymptotique.

| Level $\alpha$ | boot | asym | max sd |
|---|---|---|---|
| 1% | 0.80 | 0.80 | 0.28 |
| 5% | 5.10 | 3.60 | 0.70 |
| 10% | 10.30 | 7.00 | 0.96 |

Table 2: Niveaux empiriques (en pourcentage) dans le test qu'une variance est nulle dans un modèle non linéaire à trois effets aléatoires. L'expérience a été répétée $K = 1000$ fois $B = 300$ échantillons bootstrap ont été utilisés pour chaque expérience

Pour finir nous illustrons l'effet de la présence de paramètres de nuisance sur la procédure de test. Pour cela, on considère un modèle linéaire à 8 effets aléatoires, et nous faisons croître le nombre de variances égales à zéro. Nous comparons les niveaux empiriques estimés pour trois différentes valeurs du paramètre de seuillage.

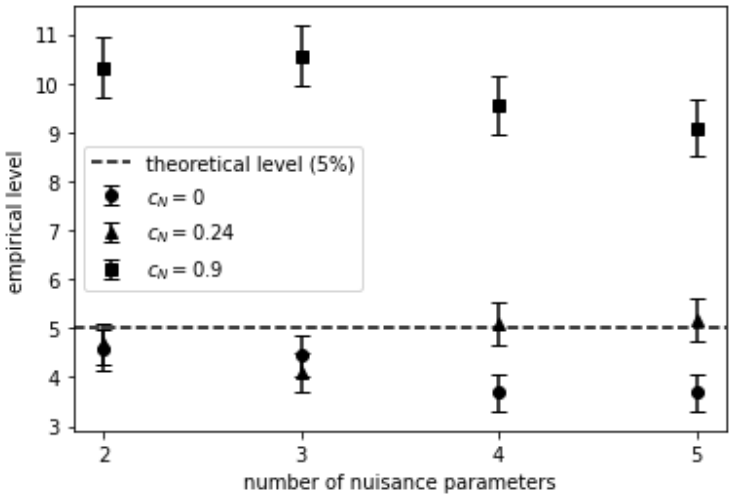



Figure 1: Comparaison des niveaux empiriques estimés pour des valeurs de seuillage différentes et un nombre de paramètres de nuisance croissant. Niveaux empiriques estimés sur $K = 2500$ jeux de données de taille $N = 30$ avec $B = 300$ échantillons bootstrap pour chaque expérience.

# 5 Conclusion

Pour conclure nous proposons une procédure de bootstrap paramétrique pour réaliser un test sur les composantes de la variance dans les modèles à effets mixtes. Notre procédure présente les avantages de 1) s'appliquer indifféremment aux modèles linéaires et non linéaires 2) prendre en compte la potentielle présence de paramètres de nuisances et 3) d'être asymptotiquement de niveau exacte, malgré les problèmes de bord et la singularité de la matrice d'information de Fisher.

# References


[1] Giuseppe Cavaliere, Heino Bohn Nielsen, Rasmus Søndergaard Pedersen, and Anders Rahbek. Bootstrap inference on the boundary of the parameter space, with application to conditional volatility models. *Journal of Econometrics*, 2020.

[2] Marie Davidian and David M Giltinan. *Nonlinear models for repeated measurement data*. Routledge, 2017.

[3] Tom Guédon, Charlotte Baey, and Estelle Kuhn. Bootstrap test procedure for variance components in nonlinear mixed effects models in the presence of nuisance parameters and singular fisher information matrix. *arXiv preprint arXiv:2306.10779*, 2023.

[4] José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2006.

[5] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

# Construction de modèles mécanistes en grande dimension avec une approche par LASSO : application à la vaccination contre le virus Ebola

Auriane Gabaut [1+] & Mélanie Prague [1*]

[1] *Université de Bordeaux, Inria, Inserm, Bordeaux Population Health Research Center, SISTM Team ; Vaccine Research Institute, Créteil, France ;*
∗ *melanie.prague@inria.fr ;* + *auriane.gabaut@inria.fr*

**Résumé.** La construction de modèles non-linéaires à effets mixtes (NLMEM) enrichit notre compréhension des processus biologiques. L'estimation dans ces modèles est facilitée par des méthodes de maximum de vraisemblance, notamment l'algorithme Stochastic Approximation Expectation-Maximization (Kuhn & Lavielle, 2005). Toutefois, cette méthode requiert une importante charge de calcul, ce qui a mené à proposer des techniques automatisées de modélisation, particulièrement pour la sélection de covariables définissant les paramètres au niveau individuel (Svensson et Jonsson, 2022 ; Aural et al.,2021). À l'instar de ces autres méthodes, en optimisant un critère d'information de type BIC, l'algorithme Stochastic Approximation for Model Building (SAMBA - Prague & Lavielle, 2022) élabore le modèle de covariables en se fondant sur la simulation de réalisations selon la loi a posteriori des paramètres individuels. Initialement, SAMBA recourt à un algorithme stepAIC pour la sélection du lien de ces covariables sur ces réalisations. Dans ce travail, nous proposons l'utilisation de la méthode LASSO pour une meilleure gestion des covariables de grande dimension. Cette méthode inclut un processus de sélection par stabilité (Meinshausen & Bühlmann, 2010). Nous avons validé notre approche à travers des simulations reproduisant la dynamique de la réponse immunitaire humorale à un vaccin contre Ebola (Pasin, 2019) en lien avec des données de transcriptomes mesurées au temps d'inclusion. Notamment, notre méthode a permis de réduire le taux de faux positifs, tout en conservant un taux de faux négatifs comparable. Nous avons mis en œuvre notre méthode avec les données de l'essai Prevac/Prevac-UP (Prevac-UP Team, 2022), comparant deux vaccins contre Ebola autorisés en Afrique.

**Mots-clés.** Apprentissage statistique ; Biostatistique ; génomique ; santé ; Grande dimension ; Sélection de modèles : Modèles non linéaires à effets mixtes.

**Abstract.**The development of nonlinear mixed-effects models (NLMEM) deepens our understanding of biological processes. Estimation in these models is facilitated by maximum likelihood methods, notably the Stochastic Approximation Expectation-Maximization algorithm (Kuhn & Lavielle, 2005). However, this method requires significant computational resources, leading to the proposal of automated modeling building techniques, especially for selecting covariates that define parameters at the individual level (SCM-Jonsson, 1998 ; COSSAC-Ayral, 2021). Similar to other models, by optimizing a Bayesian Information Criterion (BIC) type of information criterion, the Stochastic Approximation for Model Building (SAMBA-Prague & Lavielle, 2022) algorithm develops the covariate model based on realizations of the posterior distribution of individual parameters. Initially, SAMBA uses a stepAIC algorithm for the selection of the link of these covariates on these realizations. In this work, we propose the use of the LASSO methods for better management of high-dimensional covariates. This method includes a stability selection process (Meinshausen & Bühlmann, 2010). We validated our approach through simulations that replicate the dynamics of the humoral immune response to an Ebola vaccine (Pasin, 2019) linked with transcriptome data measured at the time of inclusion. Notably, our method reduced the rate of false discoveries while maintaining a comparable rate of false negatives. We implemented our method with data from the Prevac/Prevac-UP trial (Prevac-UP Team, 2022), comparing two authorized Ebola vaccines in Africa.

**Keywords.** Statistical learning ; Biostatistics ; Genomics ; Health ; High dimensionality ; Model selection : Nonlinear mixed-effects models.

# 1. Introduction

En modélisation statistique, la sélection de modèle est une étape essentielle pour identifier le modèle le plus approprié afin de représenter au mieux les données et de faire des prédictions précises. Pour faciliter ce processus, plusieurs algorithmes ont été développés. La sélection de modèle et l'utilisation d'algorithmes appropriés, tels que SCM (Stepwise Covariate Modeling) (Svensson and Jonsson, 2022), COSSAC (COnditional Sampling use for Stepwise Approach based on Correlation tests) (Ayral et al., 2021) et SAMBA (Stochastic Approximation for Model Building Algorithm) (Prague and Lavielle, 2022), sont d'une grande pertinence pour la médecine personnalisée.

## 1.1. Modèle Non-Linéaires à Effets Mixtes - Notations

Nous considérons des observations $Y_i = (Y_{ij})_{j \leq n_i}$ observée aux instants $t_{ij}$ pour l'individu $i \leq N$. Nous supposons que ces observations proviennent d'une dynamique individuelle $y_i = f(\cdot, \psi_i)$ dépendant du temps et des paramètres du modèle $\psi_i = (\psi_{il})_{l \leq m}$.

Dans le contexte des modèles non-linéaires à effets mixtes, on suppose que les paramètres individuels sont normalement distribués, $\psi_i \overset{iid}{\sim} \mathcal{N}(\psi_{pop}, \Omega)$ autour d'un paramètre de population $\psi_{pop} \in \mathbb{R}^m$. De plus, si l'individu $i$ est caractérisé par $n$ covariables $X_i = (X_{i1}, ..., X_{in})$, on peut inclure des effets de covariables dans la définition des paramètres individuels, de sorte que

$$\forall i \leq N,\ h(\psi_i) = h(\psi_{pop}) + X_i\beta + \eta_i \tag{1}$$

avec $\beta = (\beta_1, ..., \beta_m)$, où $\beta_l = (\beta_{l1}, ..., \beta_{ln})^T$ est l'effet des $n$ covariables sur le paramètre $l \leq m$, à une transformation $h$ près.

On peut finalement écrire le modèle non-linéaires à effets mixtes général comme

▸▸ **Modèle Structural** :

$$\forall i \leq N, j \leq n_i \quad \left\{ \begin{array}{rcl} y_i & = & f(\cdot, \psi_i) \\ h(\psi_i) & = & h(\psi_{pop}) + X_i\beta + \eta_i \end{array} \right. \qquad \text{(Mod Str)}$$

▸▸ **Modèle Statistique** :

$$\forall i \leq N, \quad \eta_i \overset{iid}{\sim} \mathcal{N}(0, \Omega) \qquad \text{(Mod Stat)}$$

▸▸ **Modèle d'Observation** :

$$\forall i \leq N, \quad \left\{ \begin{array}{rcl} u(Y_{ij}) & = & u(y_i(t_{ij})) + g(t_{ij}, \psi_i, \xi)\varepsilon_{ij} \\ (\varepsilon_{ij})_{j \leq n_i} & \overset{iid}{\sim} & \mathcal{N}(0, I_{n_i}) \end{array} \right. \qquad \text{(Mod Obs)}$$

(Mod)

Il est important de noter qu'il est possible de définir seulement un sous-ensemble des paramètres comme individuels, pas nécessairement tous les paramètres du modèle.

## 1.2. SAMBA

L'algorithme SAMBA (Prague and Lavielle, 2022) est itératif, il commence avec un modèle initial, qui est généralement vide $\mathscr{M}_0$ - ce qui signifie qu'il ne contient aucun effet de covariable et aucune corrélation. Ensuite, à chaque itération $k$, l'algorithme construit un nouveau modèle $\mathscr{M}_{k+1}$ basé sur le précédent $\mathscr{M}_k$. Pour ce faire, il estime les paramètres $\theta^{(k)} = \left(\psi_{pop}, \Omega, \beta, \xi\right)$ du modèle $\mathscr{M}_k$ par maximum de vraisemblance, puis échantillonne $\psi_i^{(k)}$ pour chaque individu selon la distribution à posteriori $p\left(\cdot \middle| Y, \theta^{(k)}\right)$ obtenue par Monte Carlo. À partir de ces paramètres, l'algorithme construit un modèle de covariables (ainsi qu'un modèle de correlation et d'erreur qui ne sont pas détaillés ici). Le modèle de covariable, noté $\mathscr{M}_{\mathbf{k+1}}^{\mathbf{COV}}$, se base sur les paramètres générés $\psi_i^{(k)} = (\psi_{i1}^{(k)}, ..., \psi_{im}^{(k)}),\ i \leq N$, et l'estimation $\psi_{pop}^{(k)} = (\psi_{pop\ l}^{(k)})_{l \leq m}$.

Des tests statistiques excluent les covariables qui n'affectent pas significativement un paramètre. Ensuite, un modèle de régression est construit pour le paramètre $\psi_l$, $l \leq m$ :

$$h_l(\psi_{li}^{(k)}) = h_l(\psi_{lpop}^{(k)}) + c_i\beta_l + \eta_{il}^{(k)} \tag{2}$$

où $\eta_{il}^{(k)} \sim \mathcal{N}\left(0, {\omega_l^{(k)}}^2\right)$, et $c_i$ est les covariables individuelles sélectionnées par une procédure de sélection séquentiel de covariable, stepAIC, s'il y a plus de 10 covariables, ou par une recherche exhaustive parmi tous les modèles s'il y en a moins.

# 2. Méthode

La méthode présentée propose de remplacer l'algorithme stepAIC pour la sélection de covariables dans SAMBA par une sélection par Lasso. Nous ajoutons un algorithme de sélection par stabilité pour réduire le problème d'instabilité de l'estimateur Lasso et éviter la sélection de covariables non significatives. L'algorithme initialement utilisé par SAMBA permet au critère d'information de diminuer au fil des itérations. Pour reproduire ce comportement, nous incorporons une recherche parmi plusieurs modèles de covariables, visant à minimiser le critère d'information.

## 2.1. Sélection par Lasso

Nous considérons un modèle non-linéaires à effets mixtes, comme introduit dans Mod précédemment, avec $N \in \mathbb{N}^*$ individus, $n \in \mathbb{N}^*$ covariables mesurées, et un total de $m \in \mathbb{N}^*$ paramètres dans le modèle.

À chaque itération $k \in \mathbb{N}$, la procédure SAMBA effectue d'abord, pour le modèle construit précédemment $\mathcal{M}_k$, l'estimation des paramètres de population $\theta^{(k)}$ et l'échantillonnage des paramètres individuels $(\psi_i^{(k)})_{i \leq N}$. Nous sommes alors en mesure d'écrire le modèle de régression :

$$h_l(\psi_{il}^{(k)}) = h_l(\psi_{popl}^{(k)}) + X_i\beta_l + \eta_{il}^{(k)} \tag{2}$$

Dans la nouvelle méthode, nous proposons d'effectuer la sélection par Lasso sur le modèle de régression pour chaque paramètre $l \leq m$ :

$$Y_l^{(k)} = \left(h_l\left(\psi_{il}^{(k)}\right)\right)_{i \leq N} = \left(h_l\left(\psi_{popl}^{(k)}\right)\right)_{l \leq m} + X\beta_l + \eta_l^{(k)} \tag{3}$$

avec $X = (X_1, ..., X_i)^T \in \mathcal{M}_{Nn}(\mathbb{R})$ ; $Y^{(k)}$, $Y_{pop}^{(k)} \in \mathbb{R}^m$ ; $\eta_l^{(k)} \in \mathbb{R}^m$ ; $\beta_l \in \mathbb{R}^n$.

### 2.1.1. Algorithme de sélection par stabilité

L'algorithme de sélection par stabilité (Meinshausen and Bühlmann, 2010) est une méthode développée pour rendre la sélection par Lasso plus stable et robuste. L'idée centrale de l'algorithme de sélection par stabilité est de répéter le processus de sélection par Lasso un grand nombre de fois, sur des batchs de données de taille $\lfloor N/2 \rfloor$ tirés au hasard l'ensemble des données d'observation, de taille $N \in \mathbb{N}^*$. Pour chaque batchs de données, les covariables sont sélectionnées par Lasso. Après avoir effectué ces nombreuses sélections, le nombre de fois où chaque covariable est sélectionnée dans le modèle final parmi tous les batchs est comptabilisé. Une variable est finalement conservée si sa fréquence de sélection dépasse un seuil prédéfini $t_{SS}$.

Nous proposons dans la nouvelle méthode d'utiliser soit la sélection par stabilité telle qu'elle est proposée par Meinshausen et Bühlmann, sur des batchs de données des paramètres individuels obtenus lors de l'étape de simulation ; soit de substituer les batchs de données par les réplicats obtenus lors de l'étape d'échantillonnage dans SAMBA. Pour gérer le paramètre de seuil $t_{SS}$, il est soit fixé par l'utilisateur au début de l'algorithme SAMBA, ou en le cherchant sur une grille de valeurs de 60% à 95% de sorte de minimiser le critère d'information. Nous proposons également l'utilisation de l'algorithme de calibration automatique sharp (Bodinier et al., 2023).

# 3. Illustration

Nous proposons une application des méthodes sur les données recueillies lors de l'étude PREVAC-UP du consortium PREVAC (Badio et al., 2021). Cette étude vise à évaluer la sécurité et la durabilité de la réponse immunitaire de trois stratégies de vaccins contre Ebola, dont notamment la stratégie vaccinale Ad26.ZEBOV/MVA-BN-Filo sur laquelle nous nous focaliserons.

Lors de cette étude, 1400 adultes et 1401 enfants de Guinée, Libéria, Mali et Sierroa Leone, ont reçue un vaccin contre Ebola. Ils ont ensuite été surveillés : pour chaque individu, des mesures d'anticorps, comme illustré dans la figure 1, ont été prises le jour de la première dose, puis à environ 7, 14, 28, 56 jours après. La deuxième dose est administrée 56 jours après la première. Les patients ont ensuite été suivis aux jours 63, et à 3, 6 et 12 mois après l'inclusion.

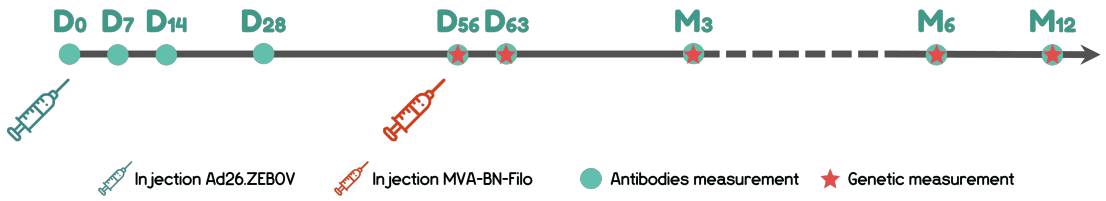



FIGURE 1 – Schéma des essais PREVAC-UP pour les personnes ayant reçu le schéma de vaccination Ad26.ZEBOV/MVA-BN-Filo.

Une sous-étude immunologique, menée spécifiquement en Guinée, rassemble 95 participants adultes, dont 30 ont reçu le schéma Ad26/MVA. Pour chacun d'eux, en plus des mesures d'anticorps, l'expression génique de plus de 28 000 gènes a été réalisée après le rappel du vaccin. Notre objectif ici est de sélectionner parmi les gènes ceux prédisant la réponse immunitaire.

La réponse immunitaire humorale a déjà été modélisée par un modèle non linéaire à effets mixtes (Pasin et al., 2019). Nous modélisons la production d'anticorps en considérant deux cellules sécrétrices d'anticorps (ASC), notées $S$ *-pour courte durée-* et $L$ *-pour longue durée-*, et caractérisées par leur demi-vie. Nous supposons que ces ASC produisent respectivement des anticorps $Ab$ aux taux $\theta_S$ et $\theta_L$. Les taux de décroissance de toutes ces entités biologiques sont respectivement notés $\delta_S$, $\delta_L$ et $\delta_{Ab}$. Ainsi, nous simulons la production d'anticorps définie par l'EDO suivante :

$$\frac{d}{dt}Ab(t) = \varphi_S e^{-\delta_S t} + \varphi_L e^{-\delta_L t} - \delta_{Ab} Ab(t)$$

Nous ajoutons ensuite des effets aléatoires sur $\delta_{Ab}, \varphi_S$ et $\varphi_L$ et testons les différents expression de gènes comme effets de covariables. Concernant le modèle d'observation, seul une mesure bruitée des niveaux d'anticorps est mesurée.

Pour illustrer les avantages de la nouvelle méthode développée par rapport à la méthode originale, nous présenterons une étude de simulation. Notre objectif est alors de tester si les méthodes proposés permettent d'obtenir un meilleur taux de faux positifs en conservant des niveaux comparables de faux négatifs parmis les covariables sélectionnés dans le modèle final.

Nous présenterons aussi les résultats de l'étude sur les données de PREVAC.

# Références

Ayral, G., Si Abdallah, J.-F., Magnard, C., and Chauvin, J. (2021). A novel method based on unbiased correlations tests for covariate selection in nonlinear mixed

effects models : The cossac approach. *CPT : Pharmacometrics & Systems Pharmacology*, 10(4) :318–329.

Badio, M., Lhomme, E., Kieh, M., Beavogui, A., Kennedy, S., Doumbia, S., Leigh, B., Sow, S., Diallo, A., Fusco, D., Kirchoff, M., Termote, M., Vatrinet, R., Wentworth, D., Esperou, H., Lane, H., Pierson, J., Watson-Jones, D., Roy, C., and Yazdanpanah, Y. (2021). Partnership for research on ebola vaccination (prevac) : protocol of a randomized, double-blind, placebo-controlled phase 2 clinical trial evaluating three vaccine strategies against ebola in healthy volunteers in four west african countries. *Trials*, 22.

Bodinier, B., Filippi, S., Nøst, T. H., Chiquet, J., and Chadeau-Hyam, M. (2023). Automated calibration for stability selection in penalised regression and graphical models. *Journal of the Royal Statistical Society Series C : Applied Statistics*, 72 :1375–1393.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, 72 :417–473.

Pasin, C., Balelli, I., Van Effelterre, T., Bockstal, V., Solforosi, L., Prague, M., Douoguih, M., and Thiébaut, R. (2019). Dynamics of the Humoral Immune Response to a Prime-Boost Ebola Vaccine : Quantification and Sources of Variation. *Journal of Virology*, 93(18) :e00579–19.

Prague, M. and Lavielle, M. (2022). Samba : A novel method for fast automatic model building in nonlinear mixed-effects models. *CPT : Pharmacometrics & Systems Pharmacology*, 11(2) :161–172.

Svensson, R. J. and Jonsson, E. N. (2022). Efficient and relevant stepwise covariate model building for pharmacometrics. *CPT : Pharmacometrics & Systems Pharmacology*, 11(9) :1210–1222.

# Méthode de classification non-paramétrique pour données longitudinales multivariées : identification de sous-phénotypes de démence

Anaïs Rouanet [1] & Carole Dufouil [2]& Cécile Proust-Lima [3]

[1] *U1219 Bordeaux Population Health research Center, Bordeaux, France, anais.rouanet@u-bordeaux.fr*
[3] *U1219 Bordeaux Population Health research Center, Bordeaux, France, carole.dufouil@u-bordeaux.fr*
[2] *U1219 Bordeaux Population Health research Center, Bordeaux, France, cecile.proust-lima@u-bordeaux.fr*

**Résumé.** De nombreuses maladies se caractérisent par des évolutions fortement hétérogènes entre patients. C'est le cas de la Maladie d'Alzheimer et des maladies apparentées (MAMA) (Reitz, 2016, Rouanet et al., 2016, Proust-Lima et al., 2016, Ten Kate et al., 2018, Eavani et al., 2018). Malgré l'abondance de biomarqueurs désormais disponibles dans les cohortes de personnes âgées pour décrire les changements pathologiques liés à la maladie d'Alzheimer (tels que la neurodégénérescence, les troubles cognitifs, l'atrophie cérébrale), cette hétérogénéité est statistiquement difficile à appréhender. Cela nécessite des méthodes de classification adaptées, capables de traiter un grand nombre de données de biomarqueurs, mesurées de manière répétée lors de visites irrégulières au fil du temps. Dans ce travail, nous avons développé un modèle de classification bayésien non paramétrique pour identifier des groupes latents de sujets à partir de marqueurs longitudinaux multivariés et de covariables transversales. L'objectif est d'identifier des sous-phénotypes latents de MAMA à partir de données de biomarqueurs répétées et de caractériser leurs voies physiopathologiques spécifiques.

L'approche de régression sur profils développée par Liverani et al. (Liverani et al., 2015) lie de manière non paramétrique une réponse longitudinale et des covariables transversales par l'intermédiaire de groupes latents. Nous avons étendu cette méthodologie à de multiples marqueurs longitudinaux. Les trajectoires de chaque marqueur sont décrites par des modèles linéaires mixtes spécifiques aux groupes, et les profils des covariables transversales sont décrits par des modèles linéaires généralisés, spécifiques aux groupes également. Un processus de Dirichlet est adopté comme a priori sur la distribution de mélange, permettant d'estimer le nombre total de groupes, et une sélection de variables basée sur une méthode de pondération est utilisée pour identifier les marqueurs qui discriminent le mieux les groupes. L'estimation des paramètres est réalisée par chaînes de Markov Monte Carlo.

Cette méthode est appliquée à la cohorte française MEMENTO (Dufouil et al., 2017) dans le but d'identifier des sous-phénotypes latents de MAMA, basés sur des tests cognitifs répétés et des volumes d'imagerie cérébrale longitudinaux et des biomarqueurs transversaux de neurodégénescence. Les résultats mettent en avant 3 sous-phénotypes de démence qui diffèrent selon la séquence et la rapidité des dégradations neuropathologiques. Chaque groupe est associé à une évolution spécifique de déclin des fonctions cognitives et à un profil spécifique d'atrophie cérébrale. En combinant l'apprentissage automatique et la modélisation biostatistique, cette approche étend les techniques de classification aux données longitudinales de

grande dimension rencontrées dans les cohortes de santé. Bien que motivée par les MAMA, elle s'applique bien au-delà de ce domaine, permettant d'identifier des profils de trajectoires.

**Mots-clés.** Données longitudinales multivariées, Classification, Bayésien, Démence

**Abstract.** Many diseases are characterized by highly heterogeneous progression patterns across patients. This is the case in Alzheimer's disease and related dementias (ADRD) (Reitz, 2016, Rouanet et al., 2016, Proust-Lima et al., 2016, Ten Kate et al., 2018, Eavani et al., 2018). Despite the abundance of biomarkers now available in ageing cohorts to describe pathological changes involved in AD (such as neurodegeneration, cognitive impairment, brain atrophy), this heterogeneity is statistically difficult to apprehend. It requires clustering methods that could handle a large number of biomarker data measured repeatedly at irregular visits over time. In this work, we developed a non-parametric Bayesian clustering model to identify latent clusters of subjects from multivariate longitudinal outcomes and additional cross-sectional variables. The objective was to uncover latent ADRD sub-phenotypes from repeated biomarker data and to characterize their specific physio pathological pathways.

We extended the profile regression approach developed by Liverani et al.(Liverani et al., 2015), that links non-parametrically a response vector and cross-sectional variables through cluster membership, to handle multiple longitudinal outcomes measured irregularly over time. The cluster-specific trajectories are described by flexible random-effect models, and the profiles of cross-sectional variables are modeled using cluster-specific generalized linear models. A Dirichlet Process prior was adopted for the mixture distribution to deal with an unconstrained number of clusters, and a variable selection based on importance weighting was used to identify markers that best discriminate between clusters. Parameter estimation is achieved using Monte Carlo Markov Chains.

This method is applied to the French MEMENTO study (Dufouil et al., 2017) to uncover latent sub-phenotypes of ADRD, based on repeated cognitive tests and cross-sectional brain imaging volumes. We identify 3 sub-phenotypes that differ according to the sequence and speed of neuropathological degradations. Each group is associated with a specific pattern of cognitive functions decline and a specific profile of brain atrophy. By combining machine learning and biostatistical modeling, this approach extends clustering techniques to large-dimensional longitudinal data encountered in health cohorts. Although motivated by ADRD, it applies far beyond as a mean to identify profiles of trajectories.

**Keywords.** Multivariate longitudinal data, clustering, Bayesian, Dementia

# 1 Introduction

Les Maladie d'Alzheimer et Maladies apparentées (MAMA) se manifestent toutes par l'apparition de la démence, qui est un syndrome caractérisé par une dégradation lente et progressive des fonctions cognitives. Les altérations physiopathologiques qui l'accompagnent ont des conséquences sur le fonctionnement social et professionnel. Cependant, elles se caractérisent par des évolutions très hétérogènes entre patients.

Les données aujourd'hui disponibles dans les cohortes sur le veillissement permettent de mieux appréhender ces altérations : la neurodégénescence est mesurée par l'accumulation de peptides amyloïd-$\beta$ ou de protéines Tau dans le cerveau, le déclin cognitif est évalué via divers tests cognitifs (mémoire, langage, attention...) et l'atrophie cérébrale est quantifiée par le volume de différentes régions dans le cerveau tel que l'hippocampe, impliqué dans le processus de mémorisation. Ces riches données permettent alors de capturer l'hétérogénéité des dégradations physiopathologiques des patients, et d'identifier des sous-phénotypes de démence, caractérisés par des séquences et des vitesses de dégradation spécifiques.

L'analyse simultanée de toutes ces données soulève ainsi plusieurs challenges : tout d'abord, il s'agit d'identifier des sous-groupes ou sous-phénotypes à partir de marqueurs longitudinaux et transversaux. Deuxièmement, le nombre de sous-phénotypes n'est pas connu a priori et doit être informé par les données. Enfin, la sélection des marqueurs clés de discrimination entre sous-phénotypes est une étape primordiale pour limiter le nombre de marqueurs, potentiellement lourds et coûteux pour les patients.

Les méthodologies proposées jusqu'alors pour analyser des marqueurs longitudinaux et transversaux dans un contexte de classification sont peu nombreuses. Liverani et al. (2015) ont proposé la régression sur profils, une méthode de classification à partir de variables transversales multiples et d'un marqueur répété. Une approche de sélection de variables permet de sélectionner les variables transversales clé pour la discrimination entre les sous-groupes. Cependant, cette méthode ne s'applique pas à de multiples marqueurs longitudinaux. Proust-Lima et al. (2017) ont proposé un modèle à classes latentes supposant l'existence d'un processus latent commun entre marqueurs. Cependant, cette hypothèse est trop forte pour analyser des marqueurs mesurant différentes dégradations pathologiques, telles que le déclin cognitif et l'atrophie cérébrale.

L'objectif de ce travail est de développer une méthode novatrice pour identifier des sous-phénotypes de démence à partir de marqueurs répétés et biomarqueurs mesurés à baseline. Pour cela, nous proposons une méthode de classification bayésienne non paramétrique pour marqueurs longitudinaux et variables transversales, incluant une méthode de sélection de variables.

# 2 Méthode

Notre méthode étend la régression sur profils, méthode bayésienne proposée par Liverani et al. (2015), à la prise en compte de marqueurs longitudinaux. Cette approche de classification se base sur un modèle de mélange infini ayant comme *a priori* un processus de Dirichlet.

Les marqueurs transversaux $W$ sont décrits par des modèles (Gaussiens dans le cas continu ou multinomiaux dans le cas discret) ayant des paramètres spécifiques aux groupes. Les marqueurs longitudinaux $Y$ sont quant à eux décrits par des modèles linéaires mixtes, également spécifiques aux groupes. Ainsi, le $m^{\text{ième}}$ marqueur $Y_{ijm}$ mesuré pour l'individu $i$, $i = 1, ..., N$, à la mesure $j$, $j = 1, ..., n_{im}$, est modélisé comme suit :

$$Y_{ijm} = X_{ij}^{(m)\top}\beta_g^{(m)} + Z_{ij}^{(m)\top}\alpha_{ig}^{(m)} + \epsilon_{ijm}$$

avec $\beta_g^{(m)}$ les paramètres de régression associés aux covariables $X_{ij}^{(m)}$, les effets aléatoires $\alpha_{ig}^{(m)} \sim \mathcal{N}(0, B^{(m)})$ associés aux covariables $Z_{ij}^{(m)}$ et les erreurs de mesure $\epsilon_{ijm} \sim \mathcal{N}(0, \sigma_m^2)$.

La vraisemblance s'écrit alors :

$$P(\theta|Y,W) = \prod_{i}^{N} \sum_{k}^{\infty} P(z_i = k;\theta) f(\mathbf{W}_i|z_i = k;\theta) f(\mathbf{Y}_i|z_i = k;\theta)$$

avec $z_i$ la variable d'appartenance aux groupes ($z_i = k$ si l'individu $i$ appartient au groupe $k$), et $\theta$ le vecteur de l'ensemble des paramètres du modèle. La méthode par pondération de sélection de variables proposée par Liverani et al. (2015) pour sélectionner les variables transversales qui influencent la partition est adaptée aux marqueurs longitudinaux. Les paramètres sont estimés par méthode de Monte Carlo par chaînes de Markov, en utilisant un échantillonneur de Gibbs.

La partition optimale de la population est définie à partir de la matrice de similarité $S$, qui représente la probabilité a posteriori de chaque paire d'individus d'être alloués au même groupe. Pour un nombre donné de groupes allant de 2 à un certain seuil, la méthode des k plus proches voisins est appliquée à la matrice de dissimilarité $(1 - S)$ pour déterminer la meilleure partition. Enfin, la partition finale est sélectionnée en maximisant le coefficient de silhouette entre toutes ces meilleures partitions.

# 3 Application

Cette méthode est appliquée à l'étude MEMENTO (Dufouil et al., 2017), cohorte clinique française de participants présentant des plaintes cognitives isolées ou un léger déficit cognitif. L'objectif est d'identifier des sous-phénotypes de démence associés à des évolutions spécifiques de déclin cognitif, d'atrophie cérébrale ainsi qu'à des profils neuropathologiques distincts.

Le déclin cognitif est quantifié par 3 tests cognitifs répétés mesurant respectivement la fluence verbale, les fonctions exécutives et la mémoire épisodique. Six biomarqueurs longitudinaux d'atrophie cérébrale mesurent le volume de 4 régions du cerveau ainsi que la glycémie et l'hyper-intensité de la substance blanche. Enfin, la neurodégénescence est mesurée à l'entrée dans l'étude par des biomarqueurs de peptite Amyloïd-$\beta$42 et de protéine Tau.

Nous identifions 3 sous-phénotypes de démence : un profil moyen de neurodégénescence, associé à un lent déclin cognitif et une lente atrophie cérébrale. Un second profil, plus jeune, caractérisé par une absence de marqueur de type Maladie d'Alzheimer. Et enfin un troisième profil biologique typique d'une maladie d'Alzheimer avec une forte neurodégénescence, un déclin cognitif marqué et une conversion rapide vers la démence.

# 4 Conclusion

En combinant l'apprentissage automatique et la modélisation biostatistique, cette approche étend les techniques de classification aux données longitudinales de grande dimension rencontrées dans les cohortes de santé. Bien que motivée par les MAMA, elle s'applique bien au-delà de ce domaine, permettant d'identifier des profils de trajectoires.

# Bibliographie


Dufouil C, Dubois B, Vellas B, et al. (2017), Cognitive and imaging markers in non-demented subjects attending a memory clinic: study design and baseline findings of the MEMENTO cohort. *Alzheimer's Research and Therapy*, 9(1):67.

Eavani H, Habes M, Satterthwaite TD, An Y, Hsieh MK, Honnorat N, Erus G, Doshi J, Ferrucci L, Beason-Held LL, Resnick SM, Davatzikos C. (2018), Heterogeneity of structural and functional imaging patterns of advanced brain aging revealed via machine learning methods. *Neurobiology of Aging*, 71:41-50.

Liverani S., Hastie D. I., Azizi L., Papathomas M., & Richardson S. (2015), PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*, 64(7), 1–30.

Proust-Lima C, Philipps V, Dartigues, J-F. (2019), A joint model for multiple dynamic processes and clinical endpoints: Application to Alzheimer's disease. *Statistics in Medicine*, 38(23):4702-4717.

Proust-Lima, C., Philipps, V., Liquet, B. (2017), Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software*, 78(2), 1–56.

Reitz, C. (2016), Toward precision medicine in Alzheimer's disease. *Annals of translational medicine*, 4(6):107.

Rouanet A, Joly P, Dartigues J-F, Proust-Lima C, Jacqmin-Gadda H. (2016), Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics*, 72(4):1123 1135.

Ten Kate M, Dicks E, Visser PJ, van der Flier WM, Teunissen CE, Barkhof F, Scheltens P, Tijms BM. (2018), Alzheimer's Disease Neuroimaging Initiative. Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline. *Brain*, 141(12):3443-3456.

# Grande dimension et parcimonie

# A variable selection method in a multivariate nonparametric regression model: Application to geoscience

Mary E. Savino [1] & Céline Lévy-Leduc [2]

[1] *Andra, 92290 Châtenay-Malabry, France and Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120 Palaiseau, France*
*mary.savino@agroparistech.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120 Palaiseau, France*
*celine.levy-leduc@agroparistech.fr*

**Résumé.** Nous présentons ici une nouvelle méthode de sélection de variables dans un modèle de régression non-paramétrique multivarié et reposant sur des données afin d'identifier les variables dont dépend réellement la fonction de régression. Cette méthode consiste à approcher la fonction sous-jacente par une combinaison linéaire de B-splines d'ordre $M$ ainsi que par la combinaison de leurs interactions deux-à-deux. Les coefficients de cette combinaison linéaire sont estimés en minimisant le critère des moindres carrés pénalisé par la somme des normes $\ell_2$ des dérivées partielles par rapport à chaque variable dont dépend la fonction. Nous montrons que la méthode proposée peut être reformulée sous la forme d'un critère de type Group Lasso. Nous validons notre approche à travers différentes expériences numériques en faisant notamment varier le nombre d'observations, le niveau de bruit et le nombre total de variables. Nous la comparons également à deux autres méthodes de l'état de l'art et une application à un système géochimique réel est présentée. A travers ces différentes applications, notre approche démontre de meilleures performances statistiques que les autres méthodes auxquelles nous l'avons comparée. Notre méthode est implémentée dans le package R `absorber` qui sera bientôt disponible sur le "Comprehensive R Archive Network" (CRAN).

**Mots-clés.** Sélection de variables, régression non paramétrique, B-splines, Group Lasso

**Summary.** In this presentation, we introduce a novel data-driven variable selection approach in a multivariate nonparametric regression model designed to capture only the variables on which the regression function depends. The underlying idea of our method consists in approximating the function by a linear combination of B-splines of order $M$ and their pairwise interactions. The coefficients of this linear combination are estimated by minimizing the penalized least-squares criterion. The penalization consists of the sum of the $\ell_2$-norms of the partial derivatives with respect to the different variables on which the function depends. We show that our proposed method can be reformulated as a Group Lasso problem. We investigate the statistical performance of our approach through numerical experiments varying the number of observations, the noise level and the total number of variables. We also compare it to two other state-of-the-art methods. An application to a geochemical system is also proposed. In these different frameworks, our approach exhibits better performance than the other methods. Our completely data-driven method is implemented in the `absorber` R package which will be soon available on the Comprehensive R Archive Network (CRAN).

**Keywords.** Variable selection, nonparametric regression, B-splines, Group Lasso

# 1 Introduction

The simulation of geochemical models that incorporate precipitation and dissolution reactions of minerals coupled to other physical processes represents a challenging task. Reactive transport modeling (RTM) serves as an illustration, striving to simultaneously consider geochemical reactions, fluid flow, heat transfer and solute transport. This challenge has led to the development of Machine Learning (ML)-based approaches aimed at estimating real solutions for full simulation models through the use of surrogate models. The main idea here consists in solving the transport equations explicitly and approximating solutions for geochemical reactions at equilibrium using surrogate models at each time step. A wealth of reviews and surveys on surrogate models for RTM is available in the works of Razavi et al. (2012); Asher et al. (2015); Jatnieks et al. (2016). Another approach to improve the surrogate model accuracy while reducing the CPU times is to reduce the number of input variables to consider in the model. This can be reformulated as a variable selection problem in the following framework. Let us consider that we have $n$ observations satisfying the following nonparametric regression model:

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \in \mathbb{R}^p, \quad 1 \leq i \leq n \tag{1}$$

where $f$ is an unknown real-valued function and where the $\varepsilon_i$'s are i.i.d centered random variables of variance $\sigma^2$. We will also assume that $f$ actually depends only on $d$ variables instead of $p$, with $d < p$, which means that there exists a real-valued function $\widetilde{f}$ such that $f(x) = \widetilde{f}(\widetilde{x})$, where $x \in \mathbb{R}^p$ and $\widetilde{x} \in \mathbb{R}^d$. Variable selection consists in identifying the components of $\widetilde{x}$.

We propose a novel method for variable selection motivated by Radchenko and James (2010) using a multivariate nonparametric regression model to retrieve the $d$ relevant variables on which $f$ in (1) depends. Our approach, presented in Section 2, consists in approximating $f$ using a linear combination of B-splines and their pairwise interactions. Additionally, drawing inspiration from the methodology of Rosasco et al. (2010), the coefficients of the linear combination are estimated by minimizing the usual least-squares criterion penalized by the sum of the $\ell_2$-norms of the partial derivatives with respect to the different variables on which $f$ depends. We show that our proposed method can be reformulated as a Group Lasso problem defined by Yuan and Lin (2006). Two different approaches to choose the penalization parameter are presented. The statistical performance of our approach are investigated in Section 3 and a geochemical application is given in Section 4.

# 2 Method

## 2.1 Approximation of $f$ using B-splines

Let $\mathbf{t}_\ell = (t_{\ell,1}, \ldots, t_{\ell,K})$ be a set of $K$ points called knots and let $\mathcal{S}_\ell$ be a compact subset of $\mathbb{R}$. Following De Boor (1978, p. 89-90) and Hastie et al. (2009, p. 160), the augmented knot sequence $\boldsymbol{\tau}_\ell$ is defined as follows:

$$\begin{aligned}
\tau_{\ell,1} = \ldots = \tau_{\ell,M} &= x_{min}^{(\ell)}, \\
\tau_{\ell,j+M} &= t_{\ell,j}, \quad j = 1, \ldots, K, \\
\tau_{\ell,K+M+1} = \ldots = \tau_{\ell,K+2M} &= x_{max}^{(\ell)}, \\
\boldsymbol{\tau}_\ell = (\tau_{\ell,1}, \ldots, \tau_{\ell,K+2M}) &= (\underbrace{x_{min}^{(\ell)}, \ldots, x_{min}^{(\ell)}}_{\text{M times}}, \underbrace{t_{\ell,1}, \ldots, t_{\ell,K}}_{\mathbf{t}_\ell}, \underbrace{x_{max}^{(\ell)}, \ldots, x_{max}^{(\ell)}}_{\text{M times}}),
\end{aligned}$$

where $x_{min}^{(\ell)}$ and $x_{max}^{(\ell)}$ are the lower and upper bounds of $\mathcal{S}_\ell$, respectively.

Denoting by $B_{k,m}^{(\ell)}$ the $k$th B-spline basis function of order $m$ with $m \leq M$ for the knot sequence $\boldsymbol{\tau}_\ell$ and for the dimension $\ell$, B-splines are defined by the following recursion:

$$B_{k,1}^{(\ell)}(x^{(\ell)}) = \begin{cases} 1 & \text{if } \tau_{\ell,k} \leq x^{(\ell)} < \tau_{\ell,k+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } \; k = 1, \ldots, K + 2M - 1, \tag{2}$$

and for $2 \leq m \leq M$,

$$B_{k,m}^{(\ell)}(x^{(\ell)}) = \frac{x^{(\ell)} - \tau_{\ell,k}}{\tau_{\ell,k+m-1} - \tau_{\ell,k}} B_{k,m-1}^{(\ell)}(x^{(\ell)}) + \frac{\tau_{\ell,k+m} - x^{(\ell)}}{\tau_{\ell,k+m} - \tau_{\ell,k+1}} B_{k+1,m-1}^{(\ell)}(x^{(\ell)}), \tag{3}$$

for $k = 1, \ldots, (K + 2M - m)$.

Inspired by Radchenko and James (2010), we propose approximating the function $f(x^{(1)}, \ldots, x^{(p)})$ appearing in (1) by a linear combination of B-splines of each variable $x^{(1)}, \ldots, x^{(p)}$ and of pairwise interaction of them as follows:

$$F(x^{(1)}, \ldots, x^{(p)}) = \sum_{\ell=1}^{p} \sum_{k=1}^{K+M} \beta_k^{(\ell)} B_k^{(\ell)}(x^{(\ell)}) + \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \left( \sum_{k=1}^{K+M} \sum_{q=1}^{K+M} \beta_{k,q}^{(\ell,j)} B_k^{(\ell)}(x^{(\ell)}) B_q^{(j)}(x^{(j)}) \right), \tag{4}$$

where $B_k^{(\ell)} = B_{k,M}^{(\ell)}$ is defined in (2) and (3) and where $\beta_k^{(\ell)}$ and $\beta_{k,q}^{(\ell,j)}$ are unknown coefficients. Observe that the column vector $(F(x_i^{(1)}, \ldots, x_i^{(p)}))_{1 \leq i \leq n}$ (4) can be rewritten as follows:

$$\sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell + \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j}. \tag{5}$$

where $\Psi_\ell$ is a $n \times (K+M)$ matrix such that its $i$th row is equal to $(B_1^{(\ell)}(x_i^{(\ell)}), \ldots, B_{K+M}^{(\ell)}(x_i^{(\ell)}))$ and $\boldsymbol{\beta}_\ell = \left(\beta_1^{(\ell)} \ \ldots \ \beta_{K+M}^{(\ell)}\right)^T$ for $1 \leq \ell \leq p$, $A^T$ denoting the transpose of the matrix $A$. Moreover, $\Phi_{\ell j}$ is an $n \times (K+M)^2$ matrix such that its $i$th row satisfies $(\Phi_{\ell j})_{i,\bullet} = ((\Psi_\ell)_{i,\bullet} \otimes (\Psi_j)_{i,\bullet})$, $\otimes$ denoting the Kronecker product, $(\Psi_\ell)_{i,\bullet}$ denoting the $i$th row of $\Psi_\ell$ and $\boldsymbol{\beta}_{\ell,j} = \left(\beta_{1,1}^{(\ell,j)} \ \beta_{1,2}^{(\ell,j)} \ \ldots \ \beta_{K+M,K+M}^{(\ell,j)}\right)^T$ for $1 \leq \ell < j \leq p$.

## 2.2 Description of our variable selection method

Inspired by the methodology of Rosasco et al. (2010), we propose selecting the variables on which $f$ depends by estimating the coefficients $\boldsymbol{\beta}_\ell$ and $\boldsymbol{\beta}_{\ell,j}$ appearing in (5) through the minimization of the following regularized criterion:

$$\begin{aligned}
&\left(\widehat{\boldsymbol{\beta}}_1(\lambda), \ldots, \widehat{\boldsymbol{\beta}}_p(\lambda), \widehat{\boldsymbol{\beta}}_{1,2}(\lambda), \ldots, \widehat{\boldsymbol{\beta}}_{(p-1),p}(\lambda)\right) \\
&= \underset{\substack{(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p) \\ (\boldsymbol{\beta}_{1,2},\ldots,\boldsymbol{\beta}_{(p-1),p})}}{\operatorname{argmin}} \left( \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j} \right\|_2^2 + \lambda \sum_{\ell=1}^{p} \sqrt{\sum_{i=1}^{n} \partial_\ell F(x_i)^2} \right),
\end{aligned}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the $Y_i$'s being defined in (1), $\partial_\ell F(x_i)$ denotes the $\ell$th partial derivative of $F$ defined in (4) at some observation point $x_i = (x_i^{(1)}, \ldots, x_i^{(p)})$ and $||y||_2^2 = \sum_{i=1}^n y_i^2$. Note that the idea underlying this criterion is that when a function does not depend on a variable its partial derivative with respect to this variable is equal to zero.

Using the definition of $F$ given in (5), the criterion can be rewritten as follows:

$$\begin{aligned}
&\left(\widehat{\boldsymbol{\beta}}_1(\lambda), \ldots, \widehat{\boldsymbol{\beta}}_p(\lambda), \widehat{\boldsymbol{\beta}}_{1,2}(\lambda), \ldots, \widehat{\boldsymbol{\beta}}_{(p-1),p}(\lambda)\right) \\
&= \underset{\substack{(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p) \\ (\boldsymbol{\beta}_{12},\ldots,\boldsymbol{\beta}_{(p-1)p})}}{\operatorname{argmin}} \left( \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j} \right\|_2^2 \right. \\
&\left. + \lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell \boldsymbol{\beta}_\ell + \sum_{j=\ell+1}^{p} (\partial_\ell \Phi_{\ell j}) \boldsymbol{\beta}_{\ell,j} + \sum_{1 \leq j < \ell} (\partial_\ell \Phi_{j\ell}) \boldsymbol{\beta}_{j,\ell} \right\|_2 \right),
\end{aligned} \tag{6}$$

where $\Psi'_\ell$ is the $n \times (K+M)$ matrix such that $(\Psi'_\ell)_{i,k} = B_k^{(\ell)'}(x_i^{(\ell)})$, $B_k^{(\ell)'}$ denoting the first derivative of $B_k^{(\ell)}$. The $i$th row of $(\partial_\ell \Phi_{\ell j})$ (resp. $(\partial_\ell \Phi_{j\ell})$) is defined by $(\partial_\ell \Phi_{\ell j})_{i,\bullet} = ((\Psi'_\ell)_{i,\bullet} \otimes (\Psi_j)_{i,\bullet})$ (resp. $(\partial_\ell \Phi_{j\ell})_{i,\bullet} = ((\Psi_j)_{i,\bullet} \otimes (\Psi'_\ell)_{i,\bullet})$). By denoting $(\partial_\ell \Phi_{\ell\bullet}) = \big((\partial_\ell \Phi_{\ell(\ell+1)}) \ldots (\partial_\ell \Phi_{\ell p})\big)$, $(\partial_\ell \Phi_{\bullet\ell}) = \big((\partial_\ell \Phi_{1\ell}) \ldots (\partial_\ell \Phi_{(\ell-1)\ell})\big)$, $\boldsymbol{\beta}_{\ell\bullet} = \left(\boldsymbol{\beta}_{\ell,(\ell+1)}^T \ldots \boldsymbol{\beta}_{\ell,p}^T\right)^T$ and $\boldsymbol{\beta}_{\bullet\ell} = \left(\boldsymbol{\beta}_{1,\ell}^T \ldots \boldsymbol{\beta}_{(\ell-1),\ell}^T\right)^T$, the penalty term can be written as:

$$\lambda \sum_{\ell=1}^{p} \left\| \Psi'_\ell \boldsymbol{\beta}_\ell + (\partial_\ell \Phi_{\ell\bullet}) \boldsymbol{\beta}_{\ell\bullet} + (\partial_\ell \Phi_{\bullet\ell}) \boldsymbol{\beta}_{\bullet\ell} \right\|_2 =: \lambda \sum_{\ell=1}^{p} \left\| (\partial_\ell \Theta_\ell) \boldsymbol{\gamma}_\ell \right\|_2, \tag{7}$$

where $\boldsymbol{\gamma}_\ell = \left(\boldsymbol{\beta}_\ell^T \ \boldsymbol{\beta}_{\ell\bullet}^T \ \boldsymbol{\beta}_{\bullet\ell}^T\right)^T$. The least-squares term can be rewritten as follows:

$$\begin{aligned}
&\left\|\mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j}\right\|_2^2 \\
&= \left\|\mathbf{Y} - \sum_{\ell=1}^{p} \Psi_\ell \boldsymbol{\beta}_\ell - \frac{1}{2}\left(\sum_{\ell=1}^{p-1} \sum_{j=\ell+1}^{p} \Phi_{\ell j} \boldsymbol{\beta}_{\ell,j} + \sum_{\ell=2}^{p} \sum_{j=1}^{\ell-1} \Phi_{j\ell} \boldsymbol{\beta}_{j,\ell}\right)\right\|_2^2 \qquad (8) \\
&=: \left\|\mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell \boldsymbol{\gamma}_\ell\right\|_2^2.
\end{aligned}$$

Equation (8) comes by setting $\Theta_1 = \begin{pmatrix}\Psi_1 & \frac{1}{2}\Phi_{1\bullet}\end{pmatrix}$ and $\Theta_p = \begin{pmatrix}\Psi_p & \frac{1}{2}\Phi_{\bullet p}\end{pmatrix}$, where $\Phi_{\ell\bullet} = (\Phi_{\ell(\ell+1)} \dots \Phi_{\ell p})$ and $\Phi_{\bullet\ell} = (\Phi_{1\ell} \dots \Phi_{(\ell-1)\ell})$. Combining (7) and (8), (6) can be rewritten as:

$$\left(\widehat{\boldsymbol{\gamma}}_1(\lambda), \dots, \widehat{\boldsymbol{\gamma}}_p(\lambda)\right) = \underset{(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p)}{\operatorname{argmin}} \left(\left\|\mathbf{Y} - \sum_{\ell=1}^{p} \Theta_\ell \boldsymbol{\gamma}_\ell\right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\|(\partial_\ell \Theta_\ell)\boldsymbol{\gamma}_\ell\right\|_2\right). \qquad (9)$$

By defining $\boldsymbol{\alpha}_\ell = (\partial_\ell \Theta_\ell)\boldsymbol{\gamma}_\ell$ and $\widetilde{\mathbf{X}}_\ell = \Theta_\ell (\partial_\ell \Theta_\ell)^+$, $A^+$ being the Moore-Penrose inverse of matrix $A$, (9) can be rewritten as:

$$(\widehat{\boldsymbol{\alpha}}_1(\lambda), \dots, \widehat{\boldsymbol{\alpha}}_p(\lambda)) = \underset{(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_p)}{\operatorname{argmin}} \left(\left\|\mathbf{Y} - \sum_{\ell=1}^{p} \widetilde{\mathbf{X}}_\ell \boldsymbol{\alpha}_\ell\right\|_2^2 + \lambda \sum_{\ell=1}^{p} \left\|\boldsymbol{\alpha}_\ell\right\|_2\right). \qquad (10)$$

The last formulation of our variable selection criterion (10) can be seen as a group lasso problem introduced by Yuan and Lin (2006), where the size $p_\ell$ of each group $\ell$ belonging to $\{1, \dots, p\}$ is equal to $n$. The coefficients $\widehat{\boldsymbol{\gamma}}_\ell(\lambda)$ are thus obtained as follows:

$$\widehat{\boldsymbol{\gamma}}_\ell(\lambda) = (\partial_\ell \Theta_\ell)^+ \widehat{\boldsymbol{\alpha}}_\ell(\lambda). \qquad (11)$$

Thus, we define the active variables for each $\lambda$ belonging to a given set $\Lambda$ as follows:

$$\mathcal{V}_\lambda = \left\{\ell, \ \sum_{k \geq 1} |\widehat{\gamma}_{\ell,k}(\lambda)| \neq 0\right\}, \qquad (12)$$

where $\widehat{\gamma}_{\ell,k}(\lambda)$ is the $k$th coefficient of $\widehat{\boldsymbol{\gamma}}_\ell(\lambda)$. We also introduce the set $\mathcal{V}_f$ of the indices of the $d$ relevant variables on which $f$ in (1) actually depends that have to be selected among the $p$ variables. We also denote the set $\overline{\mathcal{V}_f}$ of the indices of the irrelevant variables on which $f$ does not depend.

# 3 Numerical experiments

In this section, we will investigate the statistical performance of our method called ABSORBER and implemented in the `absorber` R package when the variance of the noise $\sigma^2$

increases as well as the number of observations $n$. We will also study how this novel method behaves when the number of variables $p$ grows. To demonstrate the efficiency of our method, we will compare it to two state-of-the-art methods for feature selection: LassoNet introduced in Lemhadri et al. (2021) and the widely used Random Forests (RF) introduced by Breiman (2001), using their default parameters.

## 3.1 Metrics and selection criteria

We first introduce metrics to assess the efficiency of our method:

- **True Positive Rate (TPR)** and the **False Positive Rate (FPR)**, for each $\lambda$:

$$\mathrm{TPR}(\lambda) = \frac{\mathrm{TP}(\lambda)}{d} = \frac{|\mathcal{V}_\lambda \cap \mathcal{V}_f|}{d} \quad \text{and} \quad \mathrm{FPR}(\lambda) = \frac{\mathrm{FP}(\lambda)}{p-d} = \frac{\left|\mathcal{V}_\lambda \cap \overline{\mathcal{V}_f}\right|}{p-d},$$

  where $d < p$, $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$, $\mathrm{TP}(\lambda)$ and $\mathrm{FP}(\lambda)$ are the number of true selected variables and the number of false selected variables for $\lambda$, respectively. $\mathcal{V}_\lambda$, $\mathcal{V}_f$ and $\overline{\mathcal{V}_f}$ and are introduced in (12) and in the text following this equation.

We also propose two criteria to choose $\lambda$. One allows the user to choose a threshold of percentage of selection and the other leverages the Akaike Information Criterion (AIC), introduced in Akaike (1973) to automatically choose $\lambda$. Both are defined as follows:

- **Percentage of variable selection**, for each variable $\ell$ belonging to $\{1, \ldots, p\}$:

$$P_\ell = \frac{100}{|\Lambda|} \sum_{\lambda \in \Lambda} \mathbb{1}\{\ell \in \mathcal{V}_\lambda\}, \tag{13}$$

  where $|\Lambda|$ is the total number of parameters in the set $\Lambda$, $\mathbb{1}\{A\} = 1$ if the event $A$ holds and 0 if not and $\mathcal{V}_\lambda$ is defined in (12).

- **AIC**:

$$\mathrm{AIC}(\lambda) = n \ln\left(\frac{\mathrm{RSS}(\lambda)}{n}\right) + 2T_\lambda, \tag{14}$$

  where $T_\lambda$ is the number of terms appearing in (5) by keeping only the variables selected with $\lambda$ and $\mathrm{RSS}(\lambda)$ is the residual sum of squares defined as follows:

$$\mathrm{RSS}(\lambda) = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}(\lambda)\right\|_2^2 \quad \text{with} \quad \widehat{\mathbf{Y}}(\lambda) = \sum_{\ell=1}^{p} \Theta_\ell \widehat{\boldsymbol{\gamma}}_\ell(\lambda), \tag{15}$$

  where $\widehat{\boldsymbol{\gamma}}_\ell(\lambda)$ is defined in (11). Then, the chosen $\lambda = \lambda_{\mathrm{AIC}}$ is such that:

$$\lambda_{\mathrm{AIC}} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \left(\mathrm{AIC}(\lambda)\right). \tag{16}$$

## 3.2 Results

We apply ABSORBER, LassoNet and RF on noisy observations satisfying (1) with $f = f_1$ defined as:

$$f_1(x^{(1)}, \ldots, x^{(10)}) = 1.8 \cos\left(x^{(1)}\right) \sin\left(x^{(7)} + 1\right) - 5 \ln\left(x^{(3)} + 1\right) - \frac{0.9}{{x^{(10)}}^2 + 1}, \quad (x^{(1)}, \ldots, x^{(10)}) \in [0, 1]^{10}. \tag{17}$$

and we calculate the percentage of selection defined in the previous section. Here, $\mathcal{V}_{f_1} = \{1, 3, 7, 10\}$ are the relevant variables to be selected. Note that we use $K = 1$ evenly spaced knots in the B-splines bases involved in ABSORBER. We refer the reader to Savino and Lévy-Leduc (2024) for a numerical investigation justifying this choice.

Similarly to ABSORBER, the percentage of selection can be computed for each penalization parameter of LassoNet. Concerning the RF method, we convert the percentage of increased mean square error for each variable as the model excludes them one-by-one into a percentage of selection for each of them.

The results obtained for $n = 700$ and $n = 2000$ observations with ten random samplings of the set are displayed in Figure 1. Despite being sensitive to the noise level $\sigma$ in the obser-

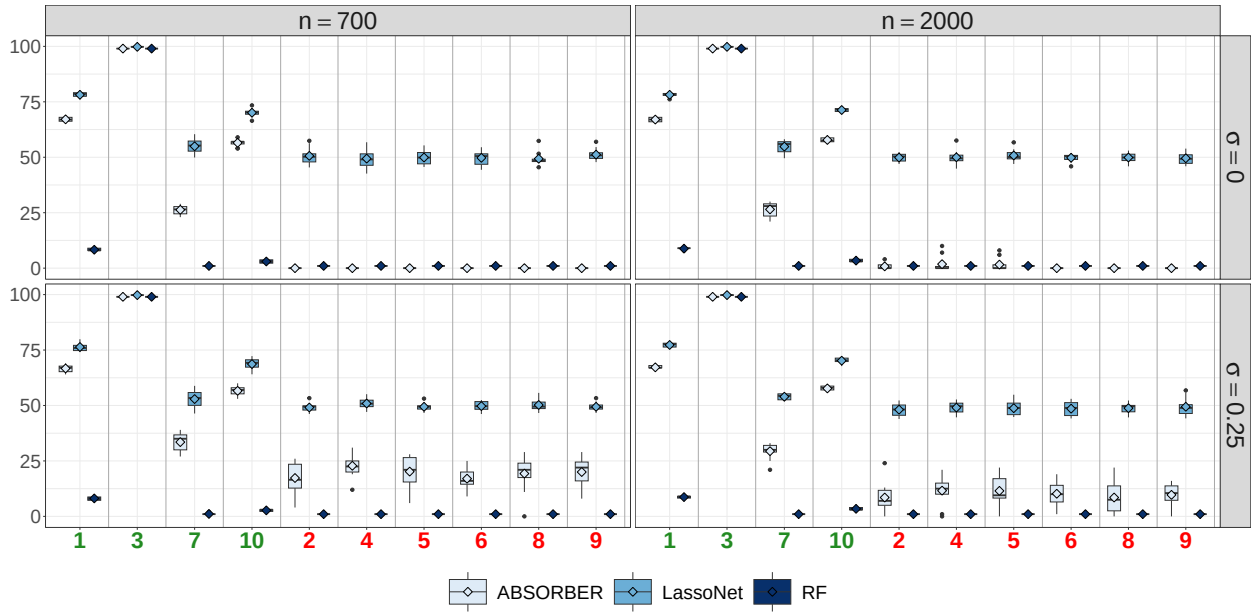



Figure 1: Percentage of selection of each variable of $f_1$ with three different methods: ABSORBER, LassoNet and RandomForests (RF) with an increasing number of observations $n$ (left to right) and of the value of $\sigma$ (top to bottom). 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

vations, ABSORBER succeeds in selecting the relevant variables with a satisfying percentage (between 25% and 40%) and allows us to choose a threshold at the visible gap (25 %) to

select the correct variables. In contrast, LassoNet and Random Forests either select irrelevant variables at a high percentage (50%) or fail to select the relevant variables, respectively. Increasing the number of observations allows us to reduce the percentage of selection for irrelevant variables with ABSORBER to 12% while maintaining the minimum percentage of relevant variables up to 30%. The two other methods appear to be unaffected by changes in $n$. These conclusions emphasize that our method outperforms those two methods for variable selection while requiring only a few parameters to choose.

We then apply our variable selection method using $\lambda_{\text{AIC}}$ as described in (16) and we calculate the value of $\text{TPR}(\lambda_{\text{AIC}})$ and $\text{FPR}(\lambda_{\text{AIC}})$. We can observe that increasing $\sigma$ slightly alters the performance of our method. However, for smaller noise levels ($\sigma < 0.25$) and with $n \geq 700$ our method enables $\text{TPR}(\lambda_{\text{AIC}}) = 1$ while maintaining $\text{FPR}(\lambda_{\text{AIC}})$ to a value smaller than 0.05. This means that almost no irrelevant variables are chosen demonstrating the efficiency of our variable selection procedure.

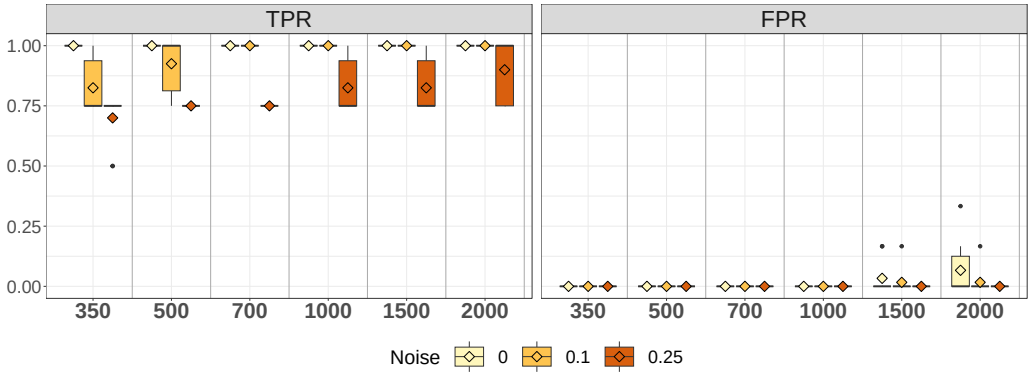



Figure 2: $\text{TPR}(\lambda)$ and $\text{FPR}(\lambda)$ values by choosing $\lambda = \lambda_{\text{AIC}}$ for $f_1$ with three noise levels in the observation sets. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

# 4 A geochemical application

We propose applying our method ABSORBER to a real geochemical system derived from the calcite dissolution and precipitation study in Kolditz et al. (2012). Our observation sets are here generated using the geochemical solver PHREEQC as in Parkhurst and Appelo (2013). For the purposes of this study, we specifically focus on the calcite precipitation/dissolution such that we can define a function $f_2$ depending on normalized concentrations and quantities of elements:

$$\begin{aligned}\text{Calcite} &= f_2(\text{C}^*, \text{Ca}^*, \text{K}^*, \text{Cl}^*, \text{Calcite}^*, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}) \\ &= \widetilde{f}_2(\text{C}^*, \text{Ca}^*, \text{Calcite}^*),\end{aligned}$$

where Calcite denotes the amount of produced calcite and C$^*$, Ca$^*$, K$^*$, Cl$^*$, Calcite$^*$ are the normalized concentrations and quantities initially present of C, Ca, K, Cl and Calcite, respectively. The normalization of each variable is done by taking into account the minimal and the maximal bound of the values so that each variable belongs to $[0, 1]$. $x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}$ are synthetic irrelevant variables obtained through uniform sampling between 0 and 1. Hereafter, $\mathcal{V}_{f_2} = \{\text{C}^*, \text{Ca}^*, \text{Calcite}^*\}$ and $\overline{\mathcal{V}_{f_2}} = \{\text{K}^*, \text{Cl}^*\, x^{(6)}, x^{(7)}, x^{(8)}, x^{(9)}, x^{(10)}\}$. We aim at retrieving $\mathcal{V}_{f_2}$ by applying our method.

The results for the application of ABSORBER, LassoNet and RF to $f_2$ are displayed in the left part of Figure 3. Our method consistently succeeds in selecting the relevant variables belonging to $\mathcal{V}_{f_2}$ (62.5% of selection) while discarding the irrelevant ones (almost 0% of selection) even with a small dataset size $n = 350$. In contrast, LassoNet selects all the variables and fails to discriminate the relevant from the irrelevant ones. Random Forests, on the other hand, tends to detect only one variable of $\mathcal{V}_{f_2}$. This shows once again that our method outperforms the other two in this geochemical case.

Furthermore, we used the AIC to select the parameter $\lambda$ and to automatically choose the relevant variables. The corresponding results are shown in right part of Figure 3. This statistical criterion proves to be highly efficient as evidenced by $\text{TPR}(\lambda_{\text{AIC}}) = 1$ and $\text{FPR}(\lambda_{\text{AIC}}) = 0$, regardless of $n$.

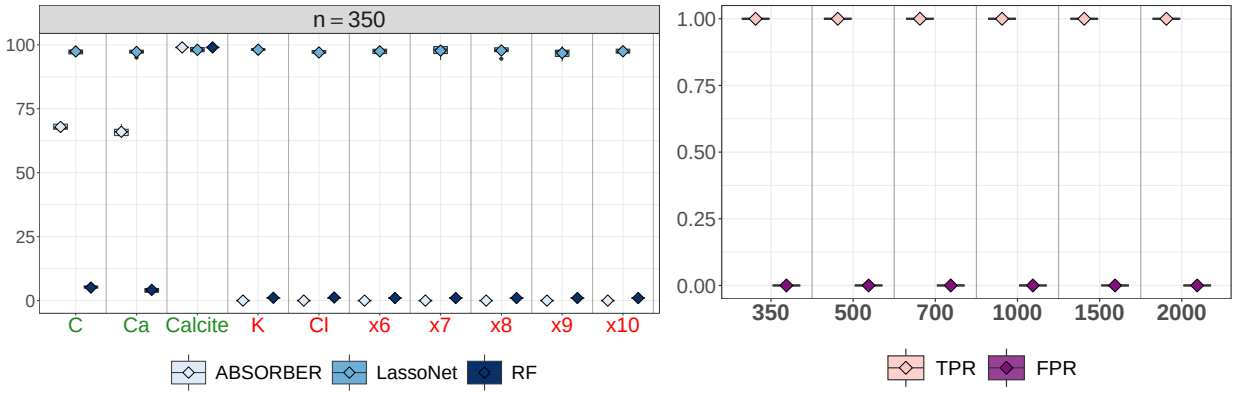



Figure 3: (Left) Percentage of selection of each variable of $f_2$ with three different methods: ABSORBER, LassoNet and RandomForests (RF) with $n = 350$ observations. (Right) $\text{TPR}(\lambda)$ and $\text{FPR}(\lambda)$ values by choosing $\lambda = \lambda_{\text{AIC}}$ for $f_2$ with an increasing number of observations $n$. 10 random samplings of $\mathbf{Y}$ were used to obtain these results. The empty diamonds inside the boxplots correspond to the mean value and the plain bullets outside the boxplots are the extreme values.

# References


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281. In B.N.Petrov, F.Csaki (Eds.).

Asher, M. J., B. F. Croke, A. J. Jakeman, and L. J. Peeters (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research 51*(8), 5957–5973.

Breiman, L. (2001). Random forests. *Machine learning 45*, 5–32.

De Boor, C. (1978). *A practical guide to splines*, Volume 27. Springer-Verlag New York.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data mining, inference, and prediction*. New York, NY, USA: Springer New York Inc.

Jatnieks, J., M. De Lucia, D. Dransch, and M. Sips (2016). Data-driven surrogate model approach for improving the performance of reactive transport simulations. *Energy Procedia 97*, 447–453.

Kolditz, O., U.-J. Görke, H. Shao, and W. Wang (2012). *Thermo-hydro-mechanical-chemical processes in porous media: benchmarks and examples*, Volume 86. Springer Science & Business Media.

Lemhadri, I., F. Ruan, L. Abraham, and R. Tibshirani (2021). Lassonet: A neural network with feature sparsity. *The Journal of Machine Learning Research 22*(1), 5633–5661.

Parkhurst, D. L. and C. Appelo (2013). *Description of input and examples for PHREEQC version 3: a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations*. U.S.G.S. Techniques and Methods.

Radchenko, P. and G. M. James (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association 105*(492), 1541–1553.

Razavi, S., B. A. Tolson, and D. H. Burn (2012). Review of surrogate modeling in water resources. *Water Resources Research 48*(7).

Rosasco, L., M. Santoro, S. Mosci, A. Verri, and S. Villa (2010). A regularization approach to nonlinear variable selection. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 653–660. JMLR Workshop and Conference Proceedings.

Savino, M. E. and C. Lévy-Leduc (2024). A novel variable selection method in a nonlinear multivariate model using B-splines with an application to geoscience. hal-04434820.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology 68*(1), 49–67.

# A scalable Bayesian method for estimating a fixed number of coordinates in the high-dimensional sparse linear regression model

Ismaël Castillo [1] & Alice L'Huillier[2] & Kolyan Ray [3] & Luke Travis [4]

[1] *LPSM, Sorbonne Université, France, ismael.castillo@upmc.fr*
[2] *LPSM, Sorbonne Université, France, alice.lhuillier@sorbonne-universite.fr*
[3] *Department of Mathematics, Imperial College London, United Kingdom, kolyan.ray@imperial.ac.uk*
[4] *Department of Mathematics, Imperial College London, United Kingdom, luke.travis15@imperial.ac.uk*

**Résumé.** Dans ce travail, on se place dans le contexte du modèle de régression linéaire en grande dimension sous contrainte de sparsité. On souhaite obtenir un intervalle de confiance pour une coordonnée (ou un nombre fixe de coordonnées) du vecteur de régression par une méthode bayésienne. Pour cela, on définit une loi a priori sur les vecteurs sparses qui cible la coordonnée d'intérêt. L'échantillonage du posterior correspondant étant coûteux, on propose une approximation variationnelle du posterior qui va, elle aussi, cibler la coordonnée d'intérêt. D'un côté, on étudie cette méthode du point de vue théorique en montrant un résultat de type Berstein-von Mises pour l'approximation variationnelle sous des hypothèses de compatibilité sur la matrice de design. Ensuite, la méthode est implémentée sur des données simulées et montre de bonnes performances dans différents cadres, tout en ayant un temps de calcul comparable aux autres méthodes fréquentistes.

**Mots-clés.** intervalle de confiance, modèle de régression linéaire en grande dimension, sparsité, approximation variationnelle, Berstein-von Mises.

**Abstract.** We study the problem of providing confidence intervals for one or a fixed number of coordinates in the high-dimensional linear regression model under sparsity constraints via a Bayesian approach. We define an appropriate sparse prior on the whole regression vector that targets the coordinates of interest. Simulating from the actual posterior distribution in this setting may be computationally intensive. To overcome this difficulty, we propose a variational approximation to the posterior designed to target the coordinates of interest. This tractable approximation can then be used to construct confidence intervals. We give theoretical guarantees for the proposed method by deriving a Berstein-von Mises theorem for the variational approximation under compatibility conditions on the design matrix. The method is implemented on simulated data illustrating that our Bayesian procedure can be computed in comparable time to other frequentist methods while exhibiting favourable performance in a number of different settings.

**Keywords.** confidence intervals, high-dimensional linear regression model, sparsity, variational approximation, Berstein-von Mises.

# 1 Introduction

In this work we consider the high-dimensional linear regression model

$$Y = X\beta^0 + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}^n$, $X$ is a given deterministic $n \times p$ design matrix with $p > n$, $\beta^0 \in \mathbb{R}^p$ is the unknown parameter, and $\varepsilon \sim \mathcal{N}_n(0, I_n)$ is a Gaussian noise. Here we are interested in the case where $\beta^0$ is sparse : only a small fraction of coefficients $\beta_i^0$ are non-zero. Our goal is to infer one coordinate or more generally a fixed number of coordinates of the regression vector $\beta^0$ by providing (asymptotic) confidence intervals.

The works [7], [5] and [3] addressed this problem by constructing confidence intervals from estimators built from the LASSO.

In a Bayesian setting, the problem of estimating one coordinate, say $\beta_1^0$, has been studied by [6]. In [6], the author proposes an appropriate model-selection type prior and derives a Berstein-von Mises (BvM) type result for the posterior induced on the first coordinate. This implies that one can construct confidence intervals for $\beta_1^0$ from the posterior by considering the credible intervals.

When using model selection type prior, as in [6], the posterior may be challenging to sample from for large dimensions $n$ and $p$ and hence the difficulty to approximate credible intervals. In this work, our goal is to develop a Bayesian method scalable to high-dimensional settings.

A popular approach to develop scalable Bayesian methods is variational Bayes where one computes the best approximation of the posterior within a class of simpler distributions. Then this (tractable) approximation is used in place of the posterior to conduct the inference. We will use this approach to develop our method as we briefly describe now.

First, we borrow from the work of [6] to define a prior that gives a posterior from which one is able to construct confidence intervals. As in [6], the posterior obtained is hard to simulate from directly. Our strategy consists in proposing a variational approximation of the posterior relying on the mean-field variational approximation studied in [4] and using it to conduct inference on $\beta_1^0$.

We derive theoretical guarantees for the proposed method. We also investigate its performance on simulated data and compare it with that of the methods proposed by [7] and [3]. Finally, we show that the method can be extended to infer a subset of $k$ coordinates. However, to be concise, we will not present this last part of our work in this abstract.

**Notations.** For $\beta \in \mathbb{R}^p$ we denote $\beta_{-1} = (\beta_i)_{i=2}^p \in \mathbb{R}^{p-1}$ the vector of the last $p-1$ coordinates of $\beta$. We denote $X_i \in \mathbb{R}^n$ the $i^{th}$ column of $X$ and $X_{-1}$ the matrix $(X_2, \ldots, X_p) \in \mathbb{R}^{n\times(p-1)}$ consisting of the last $p-1$ columns of $X$. The Lebesgue measure on $\mathbb{R}$ is denoted by $\Lambda$. The Laplace distribution on $\mathbb{R}$, denoted $\mathrm{Lap}(\mu, \sigma)$, has density proportional to $x \to e^{-|x-\mu|/\sigma}$. The Kullback-Leibler divergence between two probability distributions $P$ and $Q$ is denoted by $KL(P, Q)$.

# 2 Methodology

## Prior's construction

To build a sparse prior on $\beta \in \mathbb{R}^p$ that targets $\beta_1^0$ , we borrow two ideas used in [6] for the prior's construction. First, we put a sparse prior only on $\beta_{-1}$ instead of a sparse prior on the whole vector $\beta$ as is usually considered for estimation of the whole regression vector $\beta_0$ (see e.g. [1] [2]). Then, on the first coordinate $\beta_1$, we consider a continuous distribution. Here, we use the second idea considered in [6]. Define $\gamma_i := X_1^T X_i / \|X_1\|_2^2$ for $i = 2, \dots, p$ the rescaled correlation between the $i^{th}$ and $1^{st}$ columns of the design matrix $X$ and

$$\beta_1^* := \beta_1 + \sum_{i=2}^{p} \gamma_i \beta_i. \tag{2}$$

In the prior, we sample independently $\beta_1^*$ and $\beta_{-1}$ with $\beta_1^*$ sampled from a continuous distribution on $\mathbb{R}$ and $\beta_{-1}$ sampled from a sparse prior. As we will see in the following, this allows for a simple characterisation of the posterior. The sparse prior we consider on $\beta_{-1}$ is the model selection prior defined as follows.

**Definition 1** (Model selection prior)**.** For $d \geq 1$, $\nu$ a probability distribution on $\{0, \dots, d\}$ and $\lambda > 0$, the *model selection* prior on $u \in \mathbb{R}^d$, denoted $MS_d(\nu, \lambda)$, is defined in the following hierarchical manner:

1. The sparsity $s$ of $u$ is distributed according to $\nu$.
2. The active set $S$ of $u$, given $s$, is uniform on the $\binom{d}{s}$ subsets of $\{1, \dots, d\}$ of size $s$.
3. $u_i | S \overset{ind}{\sim} \begin{cases} \text{Lap}(\lambda), & i \in S, \\ \delta_0, & i \notin S, \end{cases}$
   for $\delta_0$ the Dirac mass at 0.

The popular spike-and-slab prior where $\beta_i \sim q \cdot \text{Lap}(\lambda) + (1 - q) \cdot \delta_0$ independently with $\delta_0$ the Dirac mass at 0 and $q \in [0, 1]$ fits within Definition 1 with $\nu \sim \text{Binomial}(d, q)$. We now define the prior we consider on the full parameter $\beta$.

**Definition 2.** For $\nu$ a probability distribution on $\{0, \dots, p-1\}$ , $\lambda > 0$ and $g$ a positive density with respect to Lebesgue measure $\Lambda$, consider the following prior distribution $\Pi$ on $\beta \in \mathbb{R}^p$:

$$\begin{aligned} \beta_{-1} &\sim MS_{p-1}(\nu, \lambda) \\ \beta_1 | \beta_{-1} &\sim g(\cdot + \sum_{i=2}^{p} \gamma_i \beta_i) d\Lambda. \end{aligned} \tag{3}$$

Note that under this prior, we indeed have $\beta_{-1}$ independent of $\beta_1^*$ (cf (2)) and $\beta_1^* \sim g$ where $g$ is a continuous distribution on $\mathbb{R}$.

We consider two particular examples of $g$ for the prior (3).

*Example* 1 (Laplace prior). For $\sigma_n > 0$, take $g = g_n$ the $\mathrm{Lap}(0, \sigma_n)$ distribution. Then the prior reduces to $\beta_{-1} \sim MS_{p-1}(\nu, \lambda)$, $\beta_1|\beta_{-1} \sim \mathrm{Lap}(-\sum_{i=2}^{p} \gamma_i \beta_i, \sigma_n)$.

*Example* 2 (Improper prior). Consider the prior measure $\Lambda \otimes MS_{p-1}(\nu, \lambda)$ which can be seen as the prior (3) with $g = 1$.

## Posterior's characterisation

Similarly as in [6], we have a simple characterisation of the posterior induced by the prior (2). Indeed, under the posterior $\beta_{-1}$ and $\beta_1^*$ are independent. The posterior distribution of $\beta_{-1}$ is the posterior distribution in a linear regression model induced by a model selection prior (Definition 1) and we have an explicit formula for the posterior density of $\beta_1^*$. This is formalized in Lemma 1.

Define $P \in \mathbb{R}^{n \times (n-1)}$ as a matrix whose columns are an orthonormal basis of $\mathrm{span}(X_1)^{\perp}$, $\check{W} = P^T X_{-1} \in \mathbb{R}^{(n-1)\times(p-1)}$ and $\check{Y} = P^T Y \in \mathbb{R}^{n-1}$.

**Lemma 1.** *Let* $\Pi$ *be the prior* (3)*. Then under the posterior distribution,* $\beta_{-1}$ *and* $\beta_1^*$ *are independent. Furthermore, the posterior distribution of* $\beta_{-1}$ *satisfies*

$$d\pi(\beta_{-1}|Y) \propto e^{-\frac{1}{2}\|\check{Y} - \check{W}\beta_{-1}\|_2^2} dMS_{p-1}(\nu, \lambda), \tag{4}$$

*where* $\check{Y} \overset{P_0}{\sim} \mathcal{N}(\check{W}\beta_{-1}^0, I_{n-1})$*. Moreover, the posterior distribution of* $\beta_1^*$ *satisfies*

$$d\pi(\beta_1^*|Y) \propto e^{-\frac{1}{2}\|X_1\|_2^2 \left(\beta_1^* - \frac{X_1^T Y}{\|X_1\|_2^2}\right)^2} g(\beta_1^*) d\beta_1^*. \tag{5}$$

*In the particular case of Example 2, the posterior is proper and* (5) *gives that* $\beta_1^*|Y \sim \mathcal{N}\left(\frac{1}{\|X_1\|_2^2} X_1^T Y, \frac{1}{\|X_1\|_2^2}\right)$.

Now that we have a characterisation of the posterior, the next step in the bayesian method is to sample from the posterior distribution of $\beta_1$ to approximate the credible intervals. From Lemma 1, one can easily derive a way of sampling from the posterior distribution of $\beta_1$: $(i)$ sample $\beta_{-1}$ according to (4), $(ii)$ sample $\beta_1^*$ according to (5), $(iii)$ compute $\beta_1 = \beta_1^* - \sum_{i=2}^{p} \gamma_i \beta_i$. Step $(ii)$ requires to sample from a 1-dimensional distribution and can be done using standard computational tools (explicit computations or MCMC) relatively quickly. However, due to the model selection prior, step $(i)$ is very intensive for large dimensions $n$ and $p$ if one uses MCMC. To overcome this difficulty, we will now consider a variational approximation of the posterior.

## Variational approximation of the posterior

The difficult part in the sampling procedure described above is to sample from the posterior distribution of $\beta_{-1}$, that is, to sample from the posterior distribution in the linear regression model when one considers the model selection prior (Definition 1). In [4], the authors study

a mean-field VB approximation of the posterior distribution in this context. They show that this approximation performs well theoretically, and also provide a fast algorithm to compute it in the specific case of the spike-and-slab prior. The results of [4] along with the decoupled form of the posterior given in Lemma 1 have inspired the following approach. The idea is to approximate the posterior distribution of $\beta_{-1}$ while retaining the exact posterior distribution of $\beta_1^*$.

More precisely, one approximates the posterior distribution of $\beta_{-1}$ by an element of the mean-field family

$$\mathcal{Q}_{-1} = \left\{ Q_{\mu,\tau,q} = \bigotimes_{i=2}^{p} q_i \mathcal{N}(\mu_i, \tau_i^2) + (1-q_i)\delta_0 : q_i \in [0,1], \mu_i \in \mathbb{R}, \tau_i \in \mathbb{R}^+ \right\},$$

obtaining

$$\hat{Q}_{-1} = \underset{Q_{-1} \in \mathcal{Q}_{-1}}{\arg\min}\, KL(Q_{-1} || \Pi_{-1}(\cdot|Y)), \tag{6}$$

for $\Pi_{-1}(\cdot|Y)$ the marginal distribution of $\beta_{-1}$ in the posterior, defined by (4). We then form a full approximation of the posterior by using the exact posterior distribution of $\beta_1^*|Y$. In other words, we consider the approximation

$$\begin{aligned} \beta_{-1} \sim \hat{Q}_{-1}, \quad \beta_1^* \sim \pi(\beta_1^*|Y), \quad \beta_{-1} \perp\!\!\!\perp \beta_1^* \\ (\beta_1, \beta_{-1}) = (\beta_1^* - \sum_{i=2}^{p} \gamma_i \beta_i, \beta_{-1}), \end{aligned} \tag{7}$$

where $\pi(\beta_1^*|Y)$ is defined in (5). In the following, we denote by $\hat{\Pi}$ the distribution on $\beta$ given by (7), and by $\hat{\Pi}(\beta_1)$ the marginal distribution of $\beta_1$ under $\hat{\Pi}$.

Then we are able to conduct inference on $\beta_1^0$ by simply plugging in the variational approximation (7) in the standard posterior-based approach. That is to say, denoting $q_\gamma$ the $\gamma$-quantile of $\hat{\Pi}(\beta_1)$, we consider the quantile-based credible interval

$$C_\gamma := (q_{\gamma/2}, q_{1-\gamma/2}), \tag{8}$$

to give a $1-\gamma$ confidence interval for $\beta_1^0$. In practice we approximate this credible interval by simulation.

# 3 Theoretical guarantees

Let us now briefly describe the theoretical guarantees that we provide for the method presented in Section 2. We show that in the asymptotic regime $n, p \to \infty$, under some conditions, the quantiles-based credible interval (8) computed from the variational distribution $\hat{\Pi}(\beta_1)$ is an asymptotic confidence interval for the truth $\beta_1^0$. Thus the variational posterior-based inference is not only computable but also reliable.

This theoretical guarantee is provided by deriving a Berstein-von Mises (BvM) type result for the variational approximation $\hat{\Pi}(\beta_1)$. More precisely, defining $\hat{\beta}_1 := \beta_1^0 + \frac{X_1^T \varepsilon}{\|X_1\|_2^2}$, we show that, under some conditions, $\hat{\Pi}(\beta_1)$ resembles a Gaussian centered at $\hat{\beta}_1$ with variance $1/\|X_1\|_2^2$.

This means formally that for $d_{BL}$ the bounded Lipschitz distance between probability distributions, we have the following result :

**Theorem 1.** *Under some conditions on the prior* (3)*, the design matrix* $X$ *and the truth* $\beta^0$*, we have*

$$d_{BL}\left(\|X_1\|_2(\hat{\Pi}(\beta_1) - \hat{\beta}_1), \mathcal{N}(0,1)\right) \xrightarrow{P_0} 0. \tag{9}$$

The distributional approximation (9) implies that the quantiles-based credible interval (8) is an asymptotic confidence interval for the truth $\beta_1^0$.

Let us now describe informally how we will prove (9) and discuss briefly the conditions we will require on the prior, the truth and the design matrix.

**Step 1 :** By the definition (7), under the variational approximation, we have $\beta_1^* \sim \pi(\beta_1^*|Y)$. We will require $g$ to not decrease too quickly to zero to have $\beta_1^* \approx \mathcal{N}(\frac{X_1^T Y}{\|X_1\|^2}, \frac{1}{\|X_1\|_2^2})$.

**Step 2:** Then one can deduce $\beta_1^* \approx \mathcal{N}(\frac{X_1^T Y}{\|X_1\|^2}, \frac{1}{\|X_1\|_2^2}) = \mathcal{N}(\hat{\beta}_1, \frac{1}{\|X_1\|_2^2}) + \sum_{i=2}^p \gamma_i \beta_i^0$. Therefore, we have

$$\beta_1 = \beta_1^* - \sum_{i=2}^p \gamma_i \beta_i \approx \mathcal{N}(\hat{\beta}_1, \frac{1}{\|X_1\|_2^2}) - \sum_{i=2}^p \gamma_i(\beta_i - \beta_i^0).$$

Consequently, if the second term is negligible with respect to the first term, namely with respect to $1/\|X_1\|_2$, then the result holds. For this, we use the bound

$$\|X_1\|_2 |\sum_{i=2}^p \gamma_i(\beta_i - \beta_i^0)| \leq \|X_1\|_2 \max_{i=2,\dots,p} |\gamma_i| \|\beta_{-1} - \beta_{-1}^0\|_1,$$

combined with a convergence rate for $\beta_{-1}$ under the variational approximation and the assumption that $\max_{i=2,\dots,p} |\gamma_i|$ is relatively small with respect to this rate. To get a convergence rate under the variational approximation, we require some assumptions on the components of the model selection prior and compatibility conditions on the design matrix $\check{W}$ and on the truth $\beta_{-1}^0$ in the same spirit as in [4].

# 4 Empirical Results

In this section we assess the performance and behaviour of our proposed variational method with the improper choice of $g \equiv 1$. We refer to this method as I-SVB throughout, and compare its performance to that of two commonly used frequentist methods from [7] and [3]

| Scenario $(n, p, s_0, M)$ | Method | Cov. | MAE | Length | Time |
|---|---|---|---|---|---|
| $(i)$ | I-SVB | **0.946** | **0.077** | 0.405 ± 0.032 | 0.344 ± 0.032 |
| $(100, 1000, 3, \log n)$ | ZZ | 0.878 | 0.113 | 0.514 ± 0.722 | **0.132 ± 0.022** |
| | JM | 0.796 | 0.121 | 0.389 ± 0.032 | 1.242 ± 0.022 |
| $(ii)$ | I-SVB | **0.948** | 0.061 | 0.281 ± 0.022 | 0.304 ± 0.032 |
| $(200, 800, 3, \log n)$ | ZZ | 0.926 | 0.064 | 0.293 ± 0.052 | 0.224 ± 0.022 |
| | JM | 0.904 | 0.065 | 0.277 ± 0.012 | 0.741 ± 0.022 |
| $(iii)$ | I-SVB | 0.956 | **0.057** | 0.282 ± 0.022 | 0.305 ± 0.032 |
| $(200, 800, 10, \log n)$ | ZZ | 0.918 | 0.064 | 0.292 ± 0.052 | **0.223 ± 0.022** |
| | JM | 0.908 | 0.064 | 0.277 ± 0.012 | 0.739 ± 0.012 |

Table 1: Assessing the performance of the uncertainty quantification provided by each method in 4 different scenarios. The best row of each column is in bold within each scenario.

(which we will refer to as ZZ and JM respectively). We take the improper prior because we found this to dominate the other choices in practice in terms of both the quality of uncertainty quantification and computation time.

Here we present three scenarios. We parameterise each scenario by the tuple $(n, p, s_0, M)$, representing that the inference is being carried out from $n$ observations of the response, with $\beta^0 \in \mathbb{R}^p$ with sparsity $s_0$, and where the non-zero entries of $\beta^0$ are given by $M$. Furthermore, the rows of $X$ are assumed to be independently multivariate Gaussian with mean 0 and covariance $Id$. For each scenario, we simulate 500 sets of observations and for each set of observations compute a 95%−credible interval for each method: for the variational method, we make a large number of samples from the variational posterior and use the empirical quantiles; for the frequentist methods, we compute the confidence intervals directly. For each method we assess: $(i)$ the coverage (the proportion of the confidence intervals which contain the true value); $(ii)$ the mean absolute error of the centering of the confidence intervals as an estimator for the truth; $(iii)$ the mean length of the confidence intervals; and $(iv)$ the mean time for computation of the confidence intervals. The three scenarios we consider are given by the following choices: $(i)$ $(100, 1000, 3, \log n)$; $(ii)$ $(200, 800, 3, \log n)$; $(iii)$ $(200, 800, 10, \log n)$. The results are given in Table 1.

We remark that the I-SVB method delivers coverage which is approximately 95% as intended, while its MAE and length are generally better than those of the ZZ method and the JM method. This demonstrates that our method is performing uncertainty quantification better than their frequentist counterparts in these scenarios. We also investigated scenarios where the columns of $X$ present non trivial correlations and we found that our method performs well with respect to the two other methods in these settings. We remark finally here that the credible sets from the variational Bayesian methods can be computed in comparable time to their frequentist counterparts.

# References


[1] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986 – 2018, 2015.

[2] Chao Gao, Aad W. van der Vaart, and Harrison H. Zhou. A general framework for Bayes structured linear models. *The Annals of Statistics*, 48(5):2848 – 2878, 2020.

[3] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909, 2014.

[4] Kolyan Ray and Botond Szabo. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117:1–31, 11 2020.

[5] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202, 2014.

[6] Dana Yang. Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression. *Electronic Journal of Statistics*, 13(2):3082 – 3094, 2019.

[7] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):217–242, 2014.

# Estimation de vecteurs aléatoires à variation régulière avec mesure spectrale discrète via le partitionnement de variables.

Alexis Boulin [1,2]

[1] *Université Côte d'Azur, CNRS, LJAD, France, aboulin@unice.fr*
[2] *Inria, Lemon*

**Résumé.** Cette étude présente une nouvelle méthode d'estimation pour les entrées et la structure d'une matrice $A$ dans le modèle à facteurs linéaires $\mathbf{X} = A\mathbf{Z} + \mathbf{E}$. Cela est appliqué à un vecteur observable $\mathbf{X} \in \mathbb{R}^d$ avec $\mathbf{Z} \in \mathbb{R}^K$, un vecteur composé de variables aléatoires à variation régulière indépendantes, et un bruit indépendant de $\mathbf{Z}$ à queue légère $\mathbf{E} \in \mathbb{R}^d$. $\mathbf{X}$ est à variation régulière et sa mesure spectrale est alors discrète et entièrement caractérisée par la matrice $A$. Nous supposerons que chaque ligne de la matrice $A$ somme à 1 et est parcimonieuse. De plus, la valeur de $K$ n'est pas connue a priori. Le problème d'identification de la matrice $A$ à partir de sa matrice de corrélation extrémale est abordé. En présence de variables pures, qui sont des éléments de $\mathbf{X}$ liés, via $A$, à un unique facteur latent, la matrice $A$ peut être reconstruite à partir de la matrice de corrélation extrémale. Nos preuves d'identifiabilité sont constructives et ouvrent la voie à notre estimation innovante pour déterminer le nombre de facteurs $K$ et la matrice $A$ à partir de $n$ observations faiblement dépendantes sur $\mathbf{X}$.

**Mots-clés.** Théorie des valeurs extrêmes, Grande dimension, Modèles à facteurs latents, Partitionnement de variables souples.

**Abstract.** This study introduces a novel estimation method for the entries and structure of a matrix $A$ in the linear factor model $\mathbf{X} = A\mathbf{Z} + \mathbf{E}$. This is applied to an observable vector $\mathbf{X} \in \mathbb{R}^d$ with $\mathbf{Z} \in \mathbb{R}^K$, a vector composed of independently regularly varying random variables, and light-tailed independent noise $\mathbf{E} \in \mathbb{R}^d$. $\mathbf{X}$ is hence regularly varying and its spectral measure is subsequently discrete and completely characterized by the matrix $A$. Each row of the matrix $A$ is both scaled and sparse. Additionally, the value of $K$ is not known a priori. The problem of identifying the matrix $A$ from its matrix of pairwise extremal correlation is addressed. In the presence of pure variables, which are elements of $\mathbf{X}$ linked, through $A$, to a single latent factor, the matrix $A$ can be reconstructed from the extremal correlation matrix. Our proofs of identifiability are constructive and pave the way for our innovative estimation for determining the number of factors $K$ and the matrix $A$ from $n$ weakly dependent observations on $\mathbf{X}$.

**Keywords.** Extremes, High dimensional estimation, Latent model, Soft clustering, Variable clustering.

# 1 Introduction

Dans cette étude, nous souhaitons estimer la matrice d'association $d \times K$, $A$, qui peut présenter de la parcimonie et sert de paramètre pour la décomposition d'un vecteur aléatoire observable $\mathbf{X}$. Cela peut être exprimé comme

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}. \tag{1}$$

Dans cette équation, $\mathbf{Z}$ représente un vecteur aléatoire non observable de dimension $K$, servant de facteur latent sous-jacent. $E \in \mathbb{R}^d$ est un bruit aléatoire non observable, avec des entrées indépendantes. Le nombre précis de facteurs, $K$, reste non divulgué et à la fois $d$ et $K$ sont autorisés à augmenter et à être plus grands que $n$, le nombre d'observations. Pour établir les fondements de notre cadre dans la théorie des valeurs extrêmes, nous supposons que $\mathbf{Z}$ est composé de variables aléatoires asymptotiquement indépendantes (voir [4, page 192] caractérisées par un paramètre de forme $\alpha$ que nous fixerons à $\alpha = 1$. Selon cette construction, le vecteur $\mathbf{Z}$ est à variation régulière dont la mesure exponentielle (voir [3, Definition 2.1.2]) est

$$\nu_{\mathbf{Z}} = \sum_{k=1}^{K} \delta_0 \otimes \cdots \otimes \nu_{Z^{(k)}} \otimes \cdots \otimes \delta_0, \quad \nu_{Z^{(k)}}(dy) = y^{-2}dy.$$

Le vecteur de pertubation $E \in \mathbb{R}^d$ est postulé posséder une distribution avec une queue plus légère que celle des facteurs associés. De plus, il présente une indépendance complète par rapport à ces facteurs. Par conséquent, $\mathbf{X}$ est également à variation régulière, ce qui peut être décrit de manière équivalente (voir, par exemple, [3, Section 2.2]) par l'existence d'une mesure angulaire $S_{\mathbf{X}}$ où la convergence faible suivante est vraie sur la sphère unité $\Delta_{d-1}$ de $\mathbb{R}^d_+$

$$\lim_{x\to\infty} \mathbb{P}\left\{ \frac{\mathbf{X}}{\|\mathbf{X}\|} \in \cdot \mid \|\mathbf{X}\| > x \right\} = S_{\mathbf{X}}(\cdot),$$

où $S_{\mathbf{X}}$ a la représentation discrète

$$S_{\mathbf{X}}(\cdot) = w^{-1} \sum_{k=1}^{K} \|A_{\cdot k}\| \delta_{\frac{A_{\cdot k}}{\|A_{\cdot k}\|}}(\cdot), \quad w = \sum_{k=1}^{K} \|A_{\cdot k}\|, \tag{2}$$

avec $\delta_x(\cdot)$ étant la mesure de Dirac qui place une masse unitaire sur $x$ et $A_{\cdot k}$ est la $k$-ème colonne de la matrice $A$.

Dans ce travail, nous proposons un partitionnement de variables basé sur ce modèle via $A$. Dans le cadre du modèle (1), nous considérons deux composantes, à savoir $X^{(i)}$ et $X^{(j)}$, appartenant au vecteur $\mathbf{X}$, comme similaires si elles partagent une association non nulle. Cette association est établie à travers la matrice $A$, les reliant à un facteur latent commun $Z^{(a)}$. Les variables présentant cette similitude sont regroupées dans le cluster désigné par $G_a$:

$$G_a = \{j \in \{1, \ldots, d\}, : A_{ja} \neq 0\}, \quad \text{pour chaque } a \in \{1, \ldots, K\}. \tag{3}$$

Étant donné que $X^{(j)}$ peut potentiellement être lié à plusieurs facteurs latents, un cluster peut déborder sur un autre.

Il convient de noter, cependant, que la définition de A dans le modèle (1) n'est pas systématiquement identifiable sans imposer de contraintes supplémentaires. Pour remédier à cela, nous envisageons une variante du modèle (1) où chaque ligne de $A$ est mise à l'échelle. Pour être précis, nous posons l'hypothèse suivante:

**Condition (i)** $\sum_{a=1}^{K} A_{ja} = 1$.

Les poids $A_{i1}, \ldots, A_{iK}$ indiquent le degré auquel les composantes s'alignent avec chaque cluster. Cette condition divise notre modèle en partitionnements durs et souples. Néanmoins, s'appuyer uniquement sur la Condition (i) est insuffisant pour garantir l'unicité de $A$ dans le modèle (1). Nous introduisons une hypothèse additionelle qui suppose la présence d'au moins une variable pure $X^{(j)}$, parmi les composantes de $\mathbf{X}$. Ces variables pures sont associées de manière unique à un seul facteur latent et à aucun autre.

**Condition (ii)** *Pour chaque* $a \in \{1, \ldots, K\}$, *il existe au moins un indice* $j \in \{1, \ldots, d\}$ *tel que* $A_{ja} = 1$ *et* $A_{jb} = 0$, $\forall b \neq a$.

# 2 Identifiabilité

Dans cette section, nous présentons une démonstration que la matrice d'association $A$, telle que définie par le modèle (1) et soumise aux conditions (i)-(ii), est identifiable, à l'exception d'une multiplication par une matrice de permutation.

Selon la construction, le vecteur $\mathbf{Z}$ est à variation régulière, il possède une matrice de corrélation extrémale représentée par $I_K$. Par conséquent, nous déduisons que le vecteur $\mathbf{X}$ est à variation régulière, conduisant à la présence d'une matrice de corrélation extrémale notée $\mathcal{X} = [\chi(i,j)]_{i=1,\ldots,d;j=1,\ldots,d}$, où

$$\chi(i,j) = \lim_{x\to\infty} \frac{\mathbb{P}\{X^{(i)} > x, X^{(j)} > x\}}{\mathbb{P}\{X^{(i)} > x\}}.$$

Le théorème suivant est destiné à démontrer que la matrice de corrélation extrémale peut être élégamment formulée en utilisant exclusivement la matrice d'association $A$. Cependant, avant d'approfondir, nous introduisons une nouvelle opération de matrice définie sur les matrices $A \in \mathcal{M}_{p,K}(\mathbb{R})$ et $B \in \mathcal{M}_{K,q}(\mathbb{R})$.

**Définition 1** *Nous appelons* $\odot$ *l'application:*

$$\begin{aligned} \odot\colon \mathcal{M}_{p,K}(\mathbb{R}) \times \mathcal{M}_{K,q}(\mathbb{R}) &\longrightarrow \mathcal{M}_{p,q}(\mathbb{R}) \\ (a_{ik}, b_{mj}) &\mapsto c_{ij}, \end{aligned}$$

*où, en notant par* $a \wedge b := \min\{a, b\}$,

$$c_{ij} = a_{i1} \wedge b_{1j} + \cdots + c_{iK} \wedge b_{Kj}.$$

Avec tous les outils à notre disposition, nous sommes prêts à présenter le théorème fondamental suivant.

**Théorème 1** *Soit* $\boldsymbol{X}$ *défini dans* (1) *et* $A$ *satisfaisant la Condition (i). Alors* $\boldsymbol{X}$ *est à variation régulière et sa matrice de corrélation extrémale* $\mathcal{X}$ *peut être écrite comme*

$$\mathcal{X} = A \odot A^{\top},$$

*avec*

$$\chi(i,j) = \sum_{k=1}^{K} A_{ik} \wedge A_{jk}.$$

Pour toute matrice d'association $A$ qui adhère au modèle (1), nous pouvons subdiviser l'ensemble $[d] = \{1, \ldots, d\}$ en deux ensembles distincts : $I$ et son complémentaire, désigné par $J$. Nous noterons $A_I$ (resp. $A_J$), la matrice $|I| \times K$ (resp. $|J| \times K$) extraite de $A$ formant des lignes dans l'ensemble d'indice $I$ (resp. $J$). Dans chaque ligne $A_{i\cdot}$ de $A_I$, il existe précisément au moins une valeur $a \in [K]$ pour laquelle $A_{ia} = 1$. Nous attribuons le terme "ensemble de variables pures" à $I$, tandis que $J$ correspond à l' "ensemble de variables impures". Pour être plus précis, pour toute matrice donnée $A$, l'ensemble de variables pures est défini comme suit

$$I = \cup_{a=1}^{K} I_a, \quad I_a := \{i \in [d] : A_{ia} = 1, A_{ib} = 0 \; \forall b \neq a\}. \tag{4}$$

Il convient de mentionner que les ensembles $\{I_a\}_{1 \leq a \leq K}$ constituent une partition de l'ensemble de variables pures $I$.

Pour établir l'identifiabilité de la matrice $A$, notre tâche est simplifiée en se concentrant sur l'identifiabilité distincte de $A_I$ et $A_J$. En ce qui concerne la définition de $A_I$, son identifiabilité est assurée tant que la partition de l'ensemble de variables pures $I$ reste identifiable. Le cœur du défi réside dans l'identifiabilité de l'ensemble $I$ et le problème inhérent de distinguer entre $I$ et $J$, basé uniquement sur la matrice de corrélation extrémale du vecteur $\mathbf{X}$. Cela constitue l'obstacle central du problème. Le théorème 2 est le pinacle de notre méthodologie. Dans la première partie **(a)**, il offre à la fois une condition nécessaire et suffisante pour identifier $[K]$ en examinant la matrice de corrélation extrémale $\mathcal{X}$. Dans la deuxième partie **(b)**, il fournit une caractérisation nécessaire et suffisante pour identifier l'ensemble $I$ lorsque la cardinalité de $I_a$ est supérieure à un. Enfin, dans la troisième partie **(c)**, il illustre que $I$ et sa partition en sous-ensembles $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$ peuvent être identifiés avec succès. Soit

$$M_i = \max_{j \in [d] \setminus \{i\}} \chi(i,j) \tag{5}$$

désignant la plus grande valeur parmi les entrées de la ligne $i$ de la matrice $\mathcal{X}$ à l'exclusion de $\chi(i,i) = 1$. De plus, soit $S_i$ l'ensemble d'indices pour lesquels $M_i$ atteint son maximum

$$S_i = \{j \in [d] \setminus \{i\}, \; \chi(i,j) = M_i\}. \tag{6}$$

**Théorème 2** *Supposons que le modèle* (1) *et les conditions (i)-(ii) sont satisfaites. Alors :*

**(a)** *L'ensemble* $[K]$ *est une clique maximale du graphe non orienté* $G = (V, E)$ *où* $V = [d]$ *et* $(i, j) \in E$ *si* $\chi(i, j) = 0$.

**(b)** *Soit* $i \in I_a$, $a \in [K]$ *et* $|I_a| \geq 2$, *alors*

$$j \in I \iff \chi(i, j) = 1 \text{ pour tout } j \in S_i.$$

**(c)** *L'ensemble de variables pures* $I$ *peut être déterminé de manière unique à partir de* $\mathcal{X}$. *De plus, sa partition* $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$ *est unique et peut être déterminée à partir de* $\mathcal{X}$ *jusqu'à des permutations d'étiquettes.*

**Théorème 3** *Supposons que le modèle* (1) *et les conditions (i)-(ii) sont satisfaites. Alors, il existe une unique matrice* $A$, *à une permutation près, telle que* $\boldsymbol{X} = A\boldsymbol{Z} + \boldsymbol{E}$ *dans* (1). *Cela implique que les clusters souples associés* $G_a$, *pour* $1 \leq a \leq K$, *sont identifiables, à une permutation près.*

# 3 Estimation

Supposons que $(\mathbf{X}_t, t \in \mathbb{Z}) = (X_t^{(1)}, \ldots, X_t^{(d)}, t \in \mathbb{Z})$ soit un processus strictement stationnaire multivarié, et que $(\mathbf{X}_t, t = 1 \ldots, n)$ soit des données observables. Soit $m \in \{1, \ldots, n\}$ un paramètre de taille de bloc et, pour $i = 1, \ldots, k$ et $j = 1, \ldots, d$, soit $M_{m,i}^{(j)} = \max\{X_t^{(j)}, : , t \in [(i-1)m, \ldots, im]\}$ le maximum des observations du $i$ème bloc dans la $j$ème coordonnée. Pour $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$, soit

$$\begin{aligned}
&\mathbf{M}_{m,i} = (M_{m,i}^{(1)}, \ldots, M_{m,i}^{(d)}),\\
&F_m^{(j)}(x) = \mathbb{P}\{M_{m,1}^{(j)} \leq x\},\\
&\mathbf{F}_m(\mathbf{x}) = (F_m^{(1)}(x^{(1)}), \ldots F_m^{(d)}(x^{(d)})),\\
&U_{m,i}^{(j)} = F_m^{(j)}(M_{m,i}^{(j)}),\\
&\mathbf{U}_{m,i} = (U_{m,i}^{(1)}, \ldots, U_{m,i}^{(d)}).
\end{aligned}$$

Ensuite, nous supposons que les marginales de $X_1^{(1)}, \ldots, X_1^{(d)}$ sont continues. Dans ce cas, les marginales de $\mathbf{M}_{m,1}$ sont également continues et

$$C_m(\mathbf{u}) = \mathbb{P}\{U_{m,1}^{(1)} \leq \mathbf{u}\}, \quad \mathbf{u} \in [0, 1]^d,$$

est la copule unique associée à $\mathbf{M}_{m,1}$. Soit $\Delta_{d-1} = \{(w^{(1)}, \ldots, w^{(d)}) \in [0, \infty)^d : \sum_{j=1}^d w^{(j)} = 1\}$ le simplexe unité dans $\mathbb{R}^d$. Tout au long, nous travaillerons sous le mécanisme suivante de génération de données.

**Définition 2 (Mécanisme de génération de données)** *Soit* $(\boldsymbol{X}_t, t \in \mathbb{Z})$ *un processus strictement stationnaire multivarié,* $(\boldsymbol{X}_t, t = 1, \ldots, n)$ *les données observables et* $C_m$ *la copule*

*associée aux maxima composante par composante pour* $m \in \{1, \ldots, n\}$. *Il existe une copule* $C_\infty$, *et une mesure Borélienne finie* $S_{\boldsymbol{X}}$ *sur* $\Delta_{d-1}$ *comme définie par* (2) *telle que*

$$\lim_{m\to\infty} C_m(\boldsymbol{u}) = C_\infty(\boldsymbol{u}), \quad \boldsymbol{u} \in [0,1]^d, \tag{MDA}$$

*où*

$$C_\infty(\boldsymbol{u}) = \exp\left\{-L\left(-\ln(u^{(1)}), \ldots, -\ln(u^{(d)})\right)\right\}$$

*et la fonction de dépendance caudale* $L : [0,\infty)^d \to [0,\infty)$ *est décrite par*

$$L(z^{(1)}, \ldots, z^{(d)}) = \sum_{a=1}^{K} \bigvee_{j=1}^{d} A_{ja} z^{(j)}.$$

Notre procédure d'estimation s'inspire de [1] et se compose des quatre étapes suivantes :

**(a)** Estimer le nombre de clusters $K$, l'ensemble de variables pures $I$ et la partition $\mathcal{I}$ ;

**(b)** Estimer $A_I$, la sous-matrice de $A$ avec les lignes $A_{i\cdot}$ correspondant à $i \in I$ ;

**(c)** Estimer $A_J$, la sous-matrice de $A$ avec les lignes $A_{j\cdot}$ correspondant à $j \in J$ ;

**(d)** Estimer les clusters souples $\mathcal{G} = \{G_1, \ldots, G_K\}$.

Dans le contexte de notre analyse, nous devons estimer les sous-matrices, désignées par $A_I$ et $A_J$, séparément. Pour commencer avec $A_I$, nous lançons la procédure d'estimation en déterminant $[K]$, ce qui nous permet ensuite d'identifier $I$ et sa partition, désignée par $\mathcal{I} = \{I_1, \ldots, I_K\}$. Pour effectuer cette étape, nous utilisons la preuve constructive fournie par le Théorème 2, en substituant l'inconnue $\mathcal{X}$ par sa version échantillonnée $\widehat{\mathcal{X}} = [\hat{\chi}_{n,m}(i,j)]_{i,j=1,\ldots,d}$. Pour plus de détails sur cette étape, veuillez vous réferer à l'Algorithme PureVar.

**Algorithm PureVar**

**procedure** PUREVAR($\widehat{\mathcal{X}}$,$\delta$)
  Initilisation : $\mathcal{I} = \emptyset$
  Construire le graphe $G = (V, E)$ où $V = [d]$ et $(i,j) \in E$ si $\hat{\chi}_{n,m}(i,j) \leq \delta$
  Trouver une clique maximal, $\mathcal{G}$, de $G$
  **for** $i \in \mathcal{G}$ **do**
    $\hat{I}^{(i)} = \{j \in [d] \setminus \{i\} : 1 - \hat{\chi}_{n,m}(i,j) \leq \delta\}$
    $\hat{I}^{(i)} = \hat{I}^{(i)} \cup \{i\}$
    $\hat{\mathcal{I}} = \mathbf{MERGE}(\hat{I}^{(i)}, \hat{\mathcal{I}})$
  Retourne $\hat{\mathcal{I}}$ et $\hat{K}$ comme le nombre d'ensembles dans $\hat{\mathcal{I}}$

Nous continuons en estimant la matrice $A_J$, ligne par ligne. Pour expliquer notre approche, nous décrivons d'abord la structure de chaque ligne, notée $A_{j\cdot}$, dans la matrice $A_J$,

pour $j \in J$. Nous devons noter que chaque $A_{j\cdot}$ satisfait des conditions de parcimonie et $\sum_{a=1}^{K} A_{ja} = 1$, tel que stipulé par la Condition (i). Ainsi, pour chaque $i \in I_a$ avec $a \in [K]$ et $j \in J$, nous avons

$$\chi(i,j) = A_{ja}.$$

En moyennant l'affichage ci-dessus sur tous les $i \in I_a$, nous obtenons

$$\frac{1}{|I_a|} \sum_{i \in I_a} \chi(i,j) = A_{ja}.$$

Par conséquent

$$\beta^{(j)} := A_{j\cdot} = \left( \frac{1}{|I_1|} \sum_{i \in I_1} \chi(i,j), \ldots, \frac{1}{|I_K|} \sum_{i \in I_K} \chi(i,j) \right),$$

qui peut être estimé à partir des données comme suit. Pour chaque $j \in \hat{J}$, nous désignons un estimateur pour le $a$-ième élément de $\beta^{(j)}$ en utilisant une approche simple. Cet estimateur est représenté comme suit

$$\bar{\chi}^{(j)} = \left( \frac{1}{|\hat{I}_1|} \sum_{i \in \hat{I}_1} \hat{\chi}_{n,m}(i,j), \ldots, \frac{1}{|\hat{I}_K|} \sum_{i \in \hat{I}_K} \hat{\chi}_{n,m}(i,j) \right).$$

Il est important de noter que cet estimateur n'est ni parcimonieux ni un élément du simplexe unitaire. Étant donné la valeur $\bar{\chi}^{(j)}$, notre objectif est de déterminer une projection euclidienne de $\bar{\chi}^{(j)}$ qui se situe dans l'espace $\mathbb{B}_0(s) = \{\beta \in \mathbb{R}^{\hat{K}}, \sum_{j=1}^{\hat{K}} \mathbb{1}_{\{\beta^{(j)} \neq 0\}} \leq s\}$, c'est-à-dire, les vecteurs ayant au plus $s$ entrées non nulles, et le simplexe unitaire $\Delta^{\hat{K}-1} = \{\beta \in \mathbb{R}^{\hat{K}},\ \beta^{(j)} \geq 0, \sum_{j=1}^{\hat{K}} \beta^{(j)} = 1\}$ :

$$\mathcal{P}(\hat{\beta}^{(j)}) \in \underset{\beta : \beta \in \mathbb{B}_0(s) \cap \Delta \hat{K}-1}{\operatorname{argmin}} ||\beta - \bar{\chi}^{(j)}||_2. \tag{7}$$

Par conséquent, pour construire un estimateur du support, nous sélectionnons uniquement les coordonnées indexées par $a$ où $\bar{\chi}_a^{(j)}$ dépasse un seuil $\delta$. Cette sélection donne un estimateur parcimonieux pour $\beta_a^{(j)}$ comme suit

$$\bar{\beta}_a^{(j)} = \bar{\chi}_a^{(j)} \mathbb{1}_{\{\bar{\chi}_a^{(j)} > \delta\}},\ j = 1, \ldots, \hat{K}.$$

Cependant, il est essentiel de noter que l'estimateur $\bar{\beta}^{(j)}$ n'appartient pas intrinsèquement au simplexe unitaire. Pour remédier à cela, nous pouvons obtenir un estimateur alternatif, noté $\hat{\beta}^{(j)}$, en projetant $\bar{\beta}^{(j)}$ sur le simplexe unitaire dans l'espace $\hat{K}$-dimensionnel. L'opération de projection sur le simplexe unitaire est réalisée en utilisant un opérateur mathématique spécifique, défini comme suit

$$(\mathcal{P}_{\Delta_{\hat{K}-1}}(\beta))_j = [\beta^{(j)} - \tau]_+,\ \tau := \frac{1}{\rho} \left( \sum_{i=1}^{\rho} \beta^{(j)} - 1 \right),$$

pour $\rho := \max k, , \beta^{(j)} > \frac{1}{k}(\sum_{j=1}^{k} w^{(j)} - 1)$. Ainsi, en désignant $\widehat{\mathcal{S}} = \operatorname{supp}(\bar{\beta}^{(j)})$, nous obtenons

$$\hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}} = \mathcal{P}_{\Delta_{\hat{K}-1}}(\bar{\beta}^{(j)}\big|_{\widehat{\mathcal{S}}}), \quad \hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}^c} = 0. \tag{8}$$

Ensuite, nous construisons la matrice $\hat{A}_{\hat{J}}$ avec des lignes correspondant aux estimateurs $\hat{\beta}^{(j)}$ pour chaque $j \in \hat{J}$. Notre estimateur final, noté $\hat{A}$, pour la matrice $A$, est obtenu en concaténant $\hat{A}_{\hat{I}}$ et $\hat{A}_{\hat{J}}$. Les propriétés statistiques de l'estimateur final sont examinées dans la Section 4, où nous fournissons également des spécifications détaillées du paramètre de réglage nécessaire à son développement.

# 4 Garanties statistiques

Nous plongeons dans l'analyse des performances statistiques de notre estimateur, noté $\hat{A}$, qui vise à estimer $A$. En plus de cette estimation, nous considérons également sa partition associée. Dans le contexte de notre section, introduisons quelques notations et concepts. Considérons la quantité $\chi(i,j)$, qui représente la corrélation extrême entre $X^{(i)}$ et $X^{(j)}$, dans le domaine d'attraction tel que spécifié par la condition MDA. Nous définissons également un paramètre crucial noté :

$$d_m = \sup_{1 \le i < j \le d} |\chi_m(i,j) - \chi(i,j)|,$$

où $\chi_m(i,j)$ est la corrélation extrême sous-asymptotique entre $M_{m,1}^{(i)}$ et $M_{m,1}^{(j)}$. Ce paramètre caractérise le biais explicite entre le cadre sous-asymptotique et le domaine d'attraction maximal. Il quantifie essentiellement le taux de convergence du système vers son comportement asymptotique. De plus, nous introduisons le nouvel événement suivant :

$$\mathcal{E} = \mathcal{E}(\delta) := \left\{ \sup_{1 \le i < j \le d} |\hat{\chi}_{n,m}(i,j) - \chi(i,j)| \le \delta \right\}. \tag{9}$$

En prenant $c_1 > 0$ suffisamment grand et

$$\delta = d_m + c_1 \left( \sqrt{\frac{\ln(kd)}{k}} + \frac{\ln(k) \ln\ln(k) \ln(kd)}{k} \right),$$

[2, Theorem 4] garantit que $\mathcal{E}$ est vérifié avec grande probabilité :

$$\mathbb{P}(\mathcal{E}) \ge 1 - d^{-c_0},$$

pour une certaine constante positive $c_0 > 0$. Nous considérons la fonction de perte pour deux matrices $d \times K$ $A, A'$ comme

$$L_2(A, A') := \min_{P \in S_K} ||AP - A'||_{\infty,2} \tag{10}$$

où $S_K$ est le groupe de toutes les matrices de permutation $K \times K$ et

$$||A||_{\infty,2} := \max_{1 \le j \le d} ||A_{j\cdot}||_2 = \max_{1 \le j \le d} \left( \sum_{j=1}^{K} |A_{ij}|^2 \right)^{1/2};$$

pour une matrice générique $A \in \mathbb{R}^{d \times K}$. Enfin, étant donné $\delta$, nous définissons

$$J_2 = \{ j \in J : \text{pour chaque } a \in [K] \text{ avec } A_{ja} \neq 0, A_{ja} > 2\delta \}. \tag{11}$$

**Théorème 4** *Soit* $(\boldsymbol{X}_t, t \in \mathbb{Z})$ *une séquence de variables aléatoires décrit par la Définition 2 avec des coefficients de mélange fort décroissants exponentiellement. Fixons* $s = \max_{j\in[d]} ||A_{j\cdot}||_0$. *Alors, sous des conditions de signaux suffisaments forts, pour l'estimateur* $\hat{A}$, *les résultats suivants sont vérifiés.*

**(a)** *Récupération des facteurs latents :*

$$\hat{K} = K,$$

**(b)** *Une borne supérieure sur* $\hat{A}$ *:*

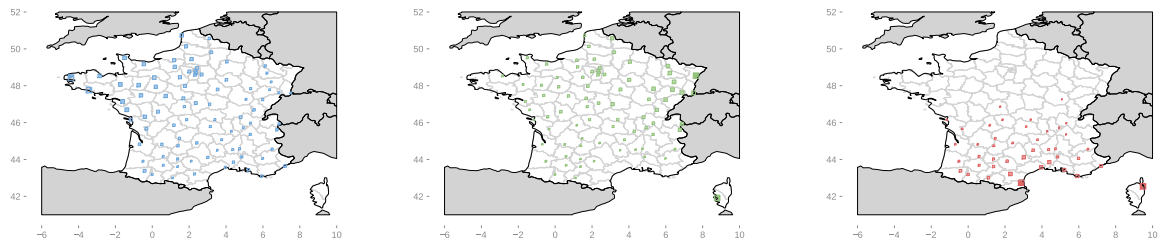
$$L_2(\hat{A}, A) \leq 4\sqrt{s}\delta,$$

**(c)** *Une garantie pour la récupération du support :*

$$supp(A_{J_2}) \subseteq supp(\hat{A}) \subseteq supp(A),$$

*avec probabilité plus grande que* $1 - d^{-c_0}$ *pour une constante positive* $c_0$.

# 5 Précipitations extrêmes en France



(a) (b) (c)

Figure 1: Dans le Panel (a), nous avons la représentation spatiale des clusters liés à la variable latente ouest. En passant au Panel (b), nous rencontrons les clusters associés à la variable latente est. Enfin, dans le Panel (c), la représentation spatiale se déploie pour les clusters liés à la variable latente sud. La force d'association de chaque emplacement avec la variable latente respective est transmise par la taille proportionnelle et l'intensité de couleur du carré.

Dans notre analyse, nous nous concentrons sur les maxima hebdomadaires des précipitations horaires enregistrées dans 92 stations météorologiques en France pendant la saison automnale, s'étendant de septembre à novembre, pour les années 1993 à 2011, ce qui donne un total de 228 maxima par bloc.

Nous proposons une approche basée sur les données pour sélectionner la valeur seuil de $\delta$. En utilisant le seuil désigné, nous révélons trois variables latentes situées dans les

régions ouest, est et sud de la France. Il est crucial de souligner que notre processus fonctionne uniquement sur la base des enregistrements de précipitations, sans aucune information géographique. Par conséquent, discerner des structures spatiales cohérentes à partir de mesures de précipitations uniquement n'est pas un résultat évident. La représentation spatiale des clusters est illustrée dans la Figure 1. La zone ouest au-dessus de Bordeaux, indiquée dans la Figure 1 Panel (a), présente des dépendances robustes avec la région centrale autour de Paris. Cependant, au-delà de ces régions, les associations avec la variable latente diminuent rapidement. Symétriquement, la région est, s'étendant de Lyon et couvrant les Vosges, l'Alsace, la Franche-Comté et les régions du nord-est de la France, représentée dans la Figure 1 Panel (b), montre des dépendances avec les régions centrales tout en diminuant rapidement en dehors de cette zone. Il est à noter que plus un emplacement est éloigné de la variable pure, moins est l'affiliation correspondante. Le cluster sud, dans la Figure 1 Panel (c), met en évidence les dépendances spatiales sur la Corse et les villes méditerranéennes. Ces associations diminuent rapidement, ce qui entraîne la formation d'un cluster moins étendu.

# References


[1] Xin Bing et al. "Adaptive estimation in structured factor models with applications to overlapping clustering". In: *The Annals of Statistics* 48.4 (2020), pp. 2055–2081. DOI: 10.1214/19-AOS1877. URL: https://doi.org/10.1214/19-AOS1877.

[2] Alexis Boulin. "Estimating Max-Stable Random Vectors with Discrete Spectral Measure using Model-Based Clustering". In: *arXiv preprint arXiv:2402.01609* (2024).

[3] Rafal Kulik and Philippe Soulier. *Heavy-tailed time series*. Springer, 2020.

[4] Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

# Procédure LASSO pour la reconstruction du support d'un processus de Hawkes multivarié en grande dimension

Christophe Denis[1,2] & Charlotte Dion-Blanc[2] & Romain E. Lacoste[1] & Laure Sansonnet[2,3]

[1] *LAMA, Université Gustave-Eiffel, France romain.lacoste@polytechnique.edu*
[2] *LPSM, Sorbonne Université, France*
[3] *MIA Paris-Saclay, Université Paris-Saclay, France*

**Résumé.** Dans cette étude, on s'intéresse au problème de la reconstruction du support de la matrice d'interaction d'un processus de Hawkes multivarié en grande dimension. Afin de composer avec la grande dimension, on impose des hypothèses de parcimonie sur la matrice d'interaction. On suppose que l'on a accès à des répétitions de trajectoires de processus de Hawkes multivariés en temps court. La stratégie proposée consiste à minimiser le contraste des moindres carrés moyenné sur les répétitions, couplé à une pénalité de type LASSO. On établit un résultat de consistance du support et de convergence de l'estimateur associé lorsque le nombre d'observations tend vers l'infini. Pour résoudre le problème de minimisation de la fonction objective, incluant un terme non-différentiable, on utilise des algorithmes de descente de gradient proximal. En présence d'une pénalité $\ell_1$, l'opérateur proximal associé s'écrit comme le seuillage doux et on utilise l'algorithme FISTA. L'implémentation de la procédure est réalisée en `C++` et bénéficie de propriétés computationelles compétitives. Enfin on propose une étude numérique sur données simulées pour valider la procédure.

**Mots-clés.** Processus de Hawkes, Grande Dimension, Parcimonie, LASSO, FISTA.

**Abstract.** In this study, we focus on the problem of support recovery of the interaction matrix of a high-dimensional multivariate Hawkes process. In order to deal with the high dimensionality, we impose sparsity assumptions on the interaction matrix. We assume that we have access to short-time path repetitions of multivariate Hawkes process. Our strategy consists in minimizing the least-squared contrast averaged over the repetitions, coupled with a LASSO-type penalty. We establish a result of support consistency and convergence of the associated estimator when the number of observations tends to infinity. To solve the problem of minimizing the objective function, which includes a non-differentiable term, we use proximal gradient descent algorithms. In the presence of a $\ell_1$ penalty, the associated proximal operator is written as a soft thresholding and the FISTA algorithm is used. The procedure is implemented in `C++` and benefits from competitive computational properties. Finally, we propose a numerical study on simulated data to validate the procedure.

**Keywords.** Hawkes Process, High Dimension, Sparsity, LASSO, FISTA.

# 1 Introduction

Les processus de Hawkes, qui furent introduits dans [10], permettent de modéliser des séries temporelles où l'occurrence d'un évènement augmente la probabilité d'en observer un nouveau dans un futur proche. En vertu de leur nature intrinsèque à modéliser des dynamiques auto-excitantes, les processus de Hawkes sont extrêmement versatiles quant à leur domaine d'applicabilité. En effet, bien qu'ils aient historiquement été utilisés pour la modélisation de secousses sismiques [11, 13], leur champ d'application s'est rapidement étendu à de nombreux autres domaines, comme les neurosciences [15], l'étude de réseaux sociaux [14], le football [2] ou encore l'écologie [6], etc.

La version multidimensionnelle de ces processus, appelée processus de Hawkes multivariés (PHM), constitue une généralisation naturelle qui enrichit considérablement les possibilités de modélisation. En effet, en plus de modéliser les interactions auto-excitantes, un tel modèle prend en compte les interactions positives entre les individus connectés au sein d'un réseau. Par conséquent, l'activité future d'une composante ne dépend plus seulement de ses activités passées, mais également de l'activité passée des individus qui interagissent avec elle et qui sont donc susceptibles d'avoir eu une influence sur elle. Là encore, les potentielles applications sont légions, comme la modélisation de potentiels d'actions au sein d'un réseau de neurone [4], la finance [8], etc.

D'un point de vue théorique, de nombreuses méthodes statistiques pour l'inférence ont été proposées pour les PHM linéaires ou non linéaires. On peut citer par exemple [1, 7, 17, 5]. En particulier dans [1], les auteurs proposent une procédure d'inférence reposant sur la minimisation du contraste des moindres carrés avec des pénalités $\ell_1$ et de faible rang, laquelle est basée sur l'observation d'une seule trajectoire d'un PHM en temps long sur l'intervalle $[0, T]$ avec $T$ qui tend vers l'infini. Dans cette étude, on propose d'adapter cette procédure à un cadre asymptotique différent. En effet, on suppose qu'on observe $n$ trajectoires d'un PHM sur l'intervalle $[0, T]$ avec $T$ fixé. Notre stratégie consiste à étudier le contraste des moindres carrés moyenné sur les répétitions avec une pénalité de LASSO. Sous des hypothèses classiques, un résultat de consistance du support et de convergence de l'estimateur associé en norme infini est établi lorsque le nombre d'observations tend vers l'infini. Enfin, pour valider la procédure, on illustre ses performances sur des données simulés.

# 2 Processus de Hawkes multivarié

Dans cette étude, on s'intéresse aux processus de Hawkes linéaires exponentiels $M$-multivariés (PHM), où $M > 1$ est la dimension du réseau. Un tel processus, noté $N = (N_1, \dots, N_M)$ est la donnée de $M$ processus ponctuels sur $\mathbb{R}_+^*$. Le processus de comptage $N_j$ est le processus $(N_j(t))_{t\in\mathbb{R}_+}$ tel que $N_j(t) = \sum_{\ell\geq 1} \mathbb{1}_{T_{j,\ell}\leq t}$, où $\{\{T_{j,\ell}\}_{\ell\geq 1}, 1 \leq j \leq M\}$ sont les temps de sauts du processus $N$. Chaque processus $N_j$ peut se caractériser par son intensité conditionnelle,

fonction définie pour tout $t \geq 0$ par :

$$\lambda_j(t) := \mu_j + \sum_{j'=1}^{M} a_{j,j'} \int_0^t \beta e^{-\beta(t-s)} \, dN_{j'}(s) = \mu_j + \sum_{j'=1}^{M} \sum_{T_{j',\ell}<t} a_{j,j'} \beta e^{-\beta(t-T_{j',\ell})}, \tag{1}$$

où

- $\mu = (\mu_1, \ldots, \mu_M) \in (\mathbb{R}_+^*)^M$ appelé vecteur d'intensité de base, régit la dynamique de chaque processus $N_j$ avant l'occurrence d'un saut et modélise l'arrivée d'évènements spontanés ;
- $A = (a_{j,j'})_{j,j'} \in \mathbb{R}_+^{M \times M}$ appelé matrice d'interaction, modélise l'influence positive de la composante $j'$ sur la composante $j$ ;
- $\beta \geq 0$ appelé taux de décroissance, contrôle les interactions entres composantes et donne la vitesse avec laquelle ces influences disparaissent avec le temps.

On dit qu'un individu $j'$ interagit avec un individu $j$ dès lors qu'il exerce une influence non-nulle sur $j$, c'est-à-dire lorsque $a_{j,j'} > 0$. Du fait de l'hypothèse de positivité des coefficients de $A$, chaque occurrence d'un évènement $T_{j',\ell}$ augmente temporairement la probabilité d'observer un nouveau saut dans un futur proche chez tous les individus avec lequel $j'$ interagit. Ainsi, un PHM permet de modéliser des effets d'excitation mutuels entre les composantes d'un réseau, lesquels dépendent des interactions survenues dans le passé.

D'après la définition (1) ci-dessus, on considère que les intensités conditionnelles des composantes du PHM $N$ dépendent d'un paramètre inconnu $\theta^*$ qui appartient à la famille de paramètres suivante :

$$\Theta := \left\{ \mu \in (\mathbb{R}_+^*)^M, \ A \in \mathcal{M}_M(\mathbb{R}_+), \ \rho(A) < 1 \right\} \in \mathbb{R}_+^{M \times M+1}$$

avec $\rho(A)$ le rayon spectral de $A$. Cette hypothèse de stabilité sous-critique assure la non-explosion du processus $N$. Dans la suite, on suppose le paramètre $\beta$ connu et commun à toutes les composantes de $N$. On note $\theta^* = (\mu^*, A^*) \in \Theta$ le paramètre à estimer et pour tout $j \in \{1, \ldots, M\}$, on note $\lambda_{j,\theta^*}$ l'intensité conditionnelle de la $j$-ème composante associée à ce paramètre. Notre objectif est de reconstruire le support de $\theta^*$, qu'on note $\text{supp}(\theta^*)$.

# 3 Cadre statistique

Soit $T > 0$ fixé. On note $\mathcal{T}_T := \{\{T_{j,\ell}\}_{1 \leq \ell \leq N_j(T)}, 1 \leq j \leq M\}$ les temps de saut d'un PHM $N = (N_1, \ldots, N_M)$ d'intensité conditionnelle $(\lambda_{1,\theta^*}, \ldots, \lambda_{M,\theta^*})$ observés en temps court sur l'intervalle $[0, T]$. On dispose d'un $n$-échantillon $D_n := \{\mathcal{T}_T^{(1)}, \ldots, \mathcal{T}_T^{(n)}\}$ qui consiste en copies indépendantes de $\mathcal{T}_T$. De plus, on se place dans le cadre de la grande dimension, à savoir qu'on considère des cas où la dimension du réseau $M$ peut être très grande (en particulier le nombre total de paramètre $M(M+1)$ peut être plus grand que $n$). Afin de composer avec la grande dimension, on impose des hypothèses de parcimonie sur la matrice d'interaction $A^*$. D'un point de vue concret, supposer $A^*$ creuse traduit que chaque individu du réseau n'est impacté que par une faible portion d'autres individus. Bien que cette hypothèse soit motivée par la réduction de la dimension du problème et pour faciliter l'interprétation, elle s'avère

très souvent naturelle du point de vue de la modélisation. À titre d'exemple, si l'on cherche à modéliser les interactions au sein de réseaux sociaux, il est vraisemblable de supposer qu'il existe des individus n'interagissant que très peu en dehors de leur cercle de connaissances proche ainsi que l'existence d'individus très peu connectés. A des fins de modélisation, on ne fait pas d'hypothèses de parcimonie sur le vecteur $\mu^*$, chaque $\mu_j^*$ étant par définition pris strictement positif. Le support de $\theta^*$, supposé de faible cardinalité, est noté $\text{supp}(\theta^*)$. Pour $j \in \{1, \dots, M\}$, on note $S_j^* := \{\theta_{j,j'}^* \neq 0,\ 1 \leq j' \leq M+1\}$ le support de la $j$-ème ligne de $\theta$. En particulier, en raison du caractère non creux de $\mu^*$, chaque $S_j^*$ contient donc au moins un élément.

On considère, à titre de fonction d'ajustement au modèle, le contraste des moindres carrés moyenné sur les répétitions défini comme suit :

$$R_{T,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{T} \sum_{j=1}^{M} \int_0^T \lambda_{j,\theta}^{(i)}(t)^2 \, dt - 2 \int_0^T \lambda_{j,\theta}^{(i)}(t) \, dN_j(t) \right), \tag{2}$$

où $\lambda_{j,\theta}^{(i)}(t)$ donné en équation (1) est définie à partir de la $i$-ème répétition.

Notre objectif est de reconstruire le support de $\theta^*$, $\text{supp}(\theta^*)$. A cette fin, étant donné l'échantillon $D_n$, notre stratégie consiste à minimiser le contraste des moindres carrés couplé à une pénalité de type LASSO :

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^{M \times M+1}}{\text{argmin}} \left\{ R_{T,n}(\theta) + \kappa \sum_{i=1}^{M} \sum_{j=1}^{M} |\theta_{i,j}| \right\}, \tag{3}$$

où $\kappa$ est la constante de régularisation à calibrer. A noter que l'on n'impose aucune contraintes de régularisation sur $\mu$.

# 4 Résultats théoriques

On commence par donner quelques notations supplémentaires. Pour une matrice $B$, on note $B'$ sa transposée. Pour tout $t \in (0, T]$, on définit la matrice aléatoire $\mathbb{H}_t \in \mathbb{R}^{n \times M+1}$ comme suit :

$$(\mathbb{H}_t)_{i,j} = H_j^{(i)}(t), \quad \text{avec} \quad H_j^{(i)}(t) := \int_0^t \beta e^{-\beta(t-s)} \, dN_j^{(i)}(s), \ j \in \{1, \dots, M\},$$

avec $H_0^{(i)} \equiv 1$. On définit aussi :

$$\mathbb{H} = \frac{1}{T} \int_0^T \mathbb{H}_t' \mathbb{H}_t \, dt.$$

Pour $j \in \{1, \dots, M\}$, on note $\mathbb{H}_{S_j^*, S_j^*} := (\mathbb{H}_{j',j'})_{j' \in S_j^*}$ la matrice extraite en supprimant les lignes et colonnes appartenant au complémentaire du support $S_j^{*c} := \{\theta_{j,j'}^* = 0,\ 1 \leq j' \leq M+1\}$.

On énonce les hypothèses suivantes qui portent sur la matrice $\mathbb{H}$.

**Hypothèse 1** (Incohérence mutuelle)**.** *Il existe un* $\gamma > 0$ *tel que*

$$\max_{j\in\{1,\dots,M\}} \|\mathbb{H}_{S_j^{*c},S_j^*}\mathbb{H}_{S_j^*,S_j^*}^{-1}\|_\infty \leq 1-\gamma$$

**Hypothèse 2** (Valeur propre minimale)**.** *Il existe* $\Lambda_0 > 0$ *tel que*

$$\min_{j\in\{1,\dots,M\}} \Lambda_{\min}\left(\frac{\mathbb{H}_{S_j^*,S_j^*}}{n}\right) \geq \Lambda_0$$

**Hypothèse 3** (Condition de signal minimal)**.**

$$\min_{j,j'\in S^*} \left|\theta^*_{j,j'}\right| > \Lambda_0 \max_j \left|S_j^*\right|^2 \frac{\log^4(nM^2)}{\sqrt{n}}$$

A présent nous pouvons énoncer le théorème principal de cette étude, on l'on établit la consistance du support et la convergence de l'estimateur associé.

**Théorème 4.** *On se place sous les hypothèses 1, 2, et 3. Soit* $\kappa = \dfrac{\log^4(nM^2)}{\sqrt{n}}$ *Pour* $n$ *assez grand, avec probabilité supérieure à* $1-\frac{C_0}{n}$ *avec* $C_0 > 0$*, le contraste des moindres carrés pénalisé défini dans l'équation* (3) *admet une unique solution* $\hat{\theta}$ *qui satisfait les propriétés suivantes :*

1. $\hat{\theta}_{j,j'} \geq 0$
2. $\mathrm{supp}(\hat{\theta}) = \mathrm{supp}(\theta^*)$ *;*
3. $\left\|\hat{\theta}-\theta^*\right\|_\infty \leq \dfrac{\Lambda_0 \max_j |S_j^*|^2 \log^4(nM^2)}{\sqrt{n}}$

En particulier, une conséquence de ce résultat est que l'on a :

$$\mathbb{P}\left(\mathrm{supp}(\hat{\theta}) = \mathrm{supp}(\theta^*)\right) \xrightarrow[n\to\infty]{} 1$$

La preuve de ce résultat suit la méthode de construction *primal-dual witness* décrite dans le Chapitre 11 de [9].

# 5 Étude numérique

On termine cette étude en proposant une étude de la procédure sur des données simulées. La simulation des trajectoires est effectuée en utilisant l'algorithme de représentation par grappe (voir [12]), qui utilise les propriétés de branchement des PHM. Pour résoudre le problème de minimisation de la fonction objective définie dans l'équation (3), incluant un terme non-différentiable, on utilise des algorithmes de descente de gradient proximal. En présence d'une pénalité $\ell_1$, l'opérateur proximal associé s'écrit comme le seuillage doux et on utilise l'algorithme FISTA (voir [3]). Le choix du pas de la descente de gradient préconisé

est $1/L$ avec $L$ la constante de Lipschitz du gradient et on peut prendre la plus grande valeur propre de la matrice hessienne. Les calculs de gradient, hessienne et d'évaluation de la fonction objective sont implémentés en `C++`. Ces derniers sont optimisés de façon à réduire considérablement le coût computationel associé. Enfin, la calibration de la constante de régularisation $\kappa$ est effectué via le critère BIC (voir [16]). Ce critère, en plus de donner de meilleurs résultats en terme de reconstruction du support, induit des temps de calculs bien moindre qu'une procédure par validation croisée. Pour toute ces raisons, prise dans son ensemble, la procédure bénéficie donc de propriétés computationelles compétitives ce qui la rend applicable à des réseaux de grande dimension.

Pour évaluer la qualité de la reconstruction du support, on utilise les deux métriques d'évaluation suivantes :

- Distance de Hamming : pour $A^*$, $\hat{A} \in \mathbb{R}_+^{M\times M}$,

$$d_H\left(A^*, \hat{A}\right) = \frac{1}{M^2}\sum_{j,j'=1}^{M} \mathbb{1}_{\left\{A^*_{j,j'} \neq \widehat{A}_{j,j'}\right\}}$$

- Distance $\ell_2$ : pour $A^*$, $\hat{A} \in \mathbb{R}_+^{M\times M}$,

$$d_{\ell_2}\left(A^*, \hat{A}\right) = \sqrt{\sum_{j,j'=1}^{M} \left|A^*_{j,j'} - \widehat{A}_{j,j'}\right|^2}$$

Afin de rendre l'étude numérique la plus complète possible, on fait varier le taux de parcimonie de la matrice d'interaction $A^*$ ainsi que sa structure, ces deux facteurs ayant un impact direct sur la difficulté de la tâche de reconstruction du support. Dans chacun de ces scénarios, on choisit $M = 25$ pour la dimension du réseau, $T = 20$ comme borne supérieure de l'intervalle d'observation et $\beta = 3$ pour le taux de décroissance. Dans les deux scénarios, on choisit $\mu^*_j = 0.5$ commun à toutes les composantes et on prend les paramètres non nuls $a^*_{j,j'}$ dans l'intervalle $[0.1, 0.5]$. Dans la Figure 1, on affiche le paramètre $\theta^*$ (où la première colonne correspond à $\mu^*$) associé à chacun des deux scénarios considérés. Dans la Figure 2, on affiche le vrai support $\mathrm{supp}(\theta^*)$ ainsi que le support $\mathrm{supp}(\hat{\theta})$ reconstruit par la procédure pour $n \in \{100, 1000, 5000\}$. Enfin dans la Table 1, on affiche les valeurs des métriques d'évaluation moyennées sur 10 répétitions Monte-Carlo dans les deux scénarios et pour les trois valeurs de $n$.

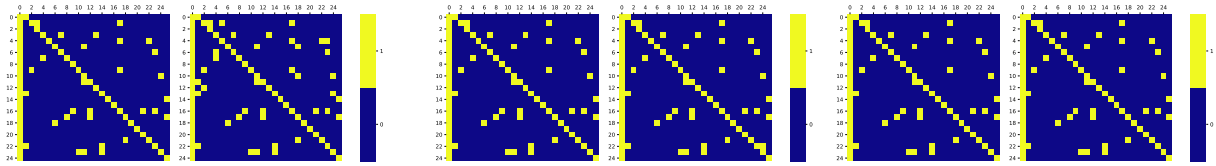
| | $d_H$ | | | $d_{l_2}$ | | |
|---|---|---|---|---|---|---|
| | $n = 100$ | $n = 1000$ | $n = 5000$ | $n = 100$ | $n = 1000$ | $n = 5000$ |
| *Scénario 1* | 0.05 (0.05) | 0.00 (0.00) | 0.00 (0.00) | 0.54 (0.11) | 0.29 (0.01) | 0.28 (0.01) |
| *Scénario 2* | 0.09 (0.05) | 0.04 (0.01) | 0.01 (0.00) | 0.53 (0.06) | 0.22 (0.01) | 0.22 (0.00) |

TABLE 1 – Valeurs des métriques obtenues moyennées sur 10 répétitions Monte-Carlo avec écart-type dans les deux scénarios et pour les trois valeurs de $n$.

Dans les deux scénarios on remarque que le support est d'autant bien reconstruit que $n$ est grand, que ce soit en terme de distance de Hamming et $\ell_2$. On peut mettre en lumière que dans le cas du scénario 2 le carré inférieur droit, ayant des valeurs $a^*_{j,j'} = 0.1$ petites, est bien détecté par la procédure.

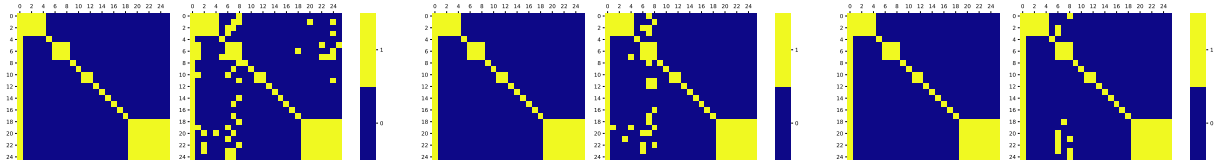
(a) *Scénario 1* (b) *Scénario 2*

FIGURE 1 – Représentation de la matrice $\theta^* = (\mu^*, A^*)$ dans les deux scénarios pour $M = 25$. Taux de parcimonie de $A^*$ dans le *Scénario 1* : 8.64%, dans le *Scénario 2* : 13.92%



(a) $n = 100$ (b) $n = 1000$ (c) $n = 5000$

FIGURE 2 – Vrai support $\text{supp}(\theta^*)$ et support reconstruit $\text{supp}(\hat{\theta})$ pour $n = 100$, pour $n = 1000$ et pour $n = 5000$ dans le *Scénario 1*.



(a) $n = 100$ (b) $n = 1000$ (c) $n = 5000$

FIGURE 3 – Vrai support $\text{supp}(\theta^*)$ et support reconstruit $\text{supp}(\hat{\theta})$ pour $n = 100$, pour $n = 1000$ et pour $n = 5000$ dans le *Scénario 2*.

# 6 Remerciements

Ce travail a été soutenu par la Chaire « Modélisation Mathématique et Biodiversité » de Veolia-École polytechnique-Museum national d'Histoire naturelle-Fondation X.

# Références

[1] E. Bacry, M. Bompaire, S. Gaïffas, and J-F Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50) :1–32, 2020.

[2] A. Baouan, S. Coustou, M. Lacome, S. Pulido, and M. Rosenbaum. Crediting football players for creating dangerous actions in an unbiased way : the generation of threat (got) indices. *arXiv preprint arXiv :2304.05242*, 2023.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1) :183–202, 2009.

[4] A. Bonnet, C. Dion-Blanc, F. Gindraud, and S. Lemler. Neuronal network inference and membrane potential model using multivariate Hawkes processes. *Journal of Neuroscience Methods*, 372 :109550, 2022.

[5] A. Bonnet, M. Martinez Herrera, and M. Sangnier. Inference of multivariate exponential Hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, 33(4) :91, 2023.

[6] C. Denis, C. Dion-Blanc, R.E. Lacoste, L. Sansonnet, and Y. Bas. Bats monitoring : A classification procedure of bats behaviors based on Hawkes processes. *hal preprint hal-04345822*, 2023.

[7] S. Donnet, V. Rivoirard, and J. Rousseau. Nonparametric Bayesian estimation for multivariate Hawkes processes. *Annals of Statistics*, 48(5) :2698–2727, 2020.

[8] P. Embrechts, T. Liniger, and L. Lin. Multivariate Hawkes processes : an application to financial data. *Journal of Applied Probability*, 48(A) :367–378, 2011.

[9] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[10] A.G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1) :83–90, 1971.

[11] A.G. Hawkes. Cluster models for earthquakes-regional comparisons. *Bulletin of the International Statistical Institute*, 45(3) :454–461, 1973.

[12] J. Møller and J.G. Rasmussen. Perfect simulation of Hawkes processes. *Advances in Applied Probability*, 37(3) :629–646, 2005.

[13] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401) :9–27, 1988.

[14] Z. Qu, C. Lyu, and C. Chi. Mush : Multi-stimuli Hawkes process based sybil attacker detector for user-review social networks. *IEEE Transactions on Network and Service Management*, 2022.

[15] P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Inference of functional connectivity in neurosciences via Hawkes processes. In *2013 IEEE global conference on signal and information processing*, pages 317–320. IEEE, 2013.

[16] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461 – 464, 1978.

[17] J. Worrall, R. Browning, P. Wu, and K. Mengersen. Fifty years later : new directions in Hawkes processes. *SORT (Statistics and Operations Research Transactions)*, 46(1) :3–38, 2022.

# Semi-LASSO : un weighted LASSO pour l'intégration de régresseurs connus dans un modèle linéaire

Anouk Rago [1] & Nicolas Champagnat [2] & Anne Gégout-Petit [3]

[1] *Université de Lorraine, CNRS, Inria, IECL, France, anouk.rago@univ-lorraine.fr*
[2] *Université de Lorraine, CNRS, Inria, IECL, France, nicolas.champagnat@inria.fr*
[3] *Université de Lorraine, CNRS, Inria, IECL, France, anne.gegout-petit@univ-lorraine.fr*

**Résumé.** Le LASSO est une technique très largement utilisée lorqu'il s'agit à la fois d'estimer les paramètres d'un modèle et d'effectuer une sélection de variables. Il est particulièrement utile pour étudier de grands jeux de données, comme cela peut être le cas en biologie des systèmes par exemple, ce qui le rend très utilisé dans le domaine de l'inférence de réseaux de gènes. Cette méthode peut par ailleurs être enrichie et améliorée par des connaissances préalables sur les régresseurs potentiels, afin de guider la sélection de variables. Dans ce cas, on peut employer un weighted LASSO, dérivé du LASSO original, dans lequel l'ajout de poids spécifiques à chaque variable permet d'encoder des a priori. Le package R 'glmnet' permet à l'utilisateur de spécifier ses propres poids via un paramètre. Dans ce papier, nous introduisons une nouvelle méthode appelée semi-LASSO qui résout un cas spécifique de weighted LASSO. Son implémentation repose sur l'utilisation du package 'glmnet', mais inclut une première étape de réduction de dimension pour une meilleure optimisation de la fonction de coût du LASSO. Des simulations numériques sont effectuées sur des données synthétiques afin de comparer les résultats obtenus avec le weighted LASSO de 'glmnet' et notre méthode semi-LASSO.

**Mots-clés.** LASSO, régression, connaissance a priori, R, réseaux de gènes

**Abstract.** The LASSO is a widely used technique when it comes to perform both estimation of parameters and variable selection. It is particularly convenient when one has to deal with large datasets like genomic data, and is thus very used in the field of gene networks inference. Nevertheless, the LASSO method can be improved thanks to prior knowledge on the potential regessors of the model. In this case, a weighted LASSO is used, the weights role is to encode the prior information. The 'glmnet' package proposes to the user to tune the 'penalty.factor' parameter to specify his own weights. We introduce in this paper a new method called semi-LASSO, which solves a specific case of weighted LASSO. Even if its implementation relies on the 'glmnet' package, it first reduces the space dimension for the LASSO optimization. Numerical experiments are performed on synthetic data to compare the performances of these two methods.

**Keywords.** LASSO, regression, prior knowledge, R, gene network inference

# 1 Introduction

Le LASSO, (*Tibshirani* 1996), est une technique très largement utilisée lorqu'il s'agit à la fois d'estimer les paramètres d'un modèle et d'effectuer une sélection de variables. Il est particulièrement utile pour étudier de grands jeux de données, comme cela peut être le cas en biologie par exemple. C'est pour cette raison qu'un grand nombre de méthodes utilisant le LASSO et ses dérivés, (*Zou* 2006), ont été développées pour inférer des réseaux de régulations de gènes,(*Sanguinetti et al* 2019), qui décrivent les relations existant entre les gènes. Cependant, reconstruire ce réseau en utilisant seulement des données transcriptomiques est une tâche complexe, notamment à cause du très grand nombre de gènes en présence ($p \approx 20\,000$), relativement au faible nombre d'échantillons $n$ recueillis : le LASSO est alors un bon outil pour sélectionner les régresseurs les plus pertinents parmi les gènes. Il n'est pas rare que des connaissances biologiques sur la structure du réseaux ou certains liens entre les protéines ou les gènes soient connus des biologistes. Lorsqu'on utilise un LASSO, une façon d'inclure ces connaissances est de spécifier différents poids pour chaque variable dans le terme de pénalisation, comme cela a été fait par exemple par *Bergensen et al* (2019). Il a de plus été montré qu'optimiser la quantité de connaissances à ajouter dans le modèle donne des résultats significativement meilleurs (*Cassan et al* 2023). D'un point de vue plus large, spécifier des pénalisations différentes peut être utile dans n'importe quel problème de régression pour lequel de l'information sur les régresseurs potentiels est disponible. Le package R 'glmnet' offre déjà une implémentation du weighted LASSO où l'utilisateur peut spécifier ses propres poids via le paramètre 'penalty.factor'.

Nous nous intéressons ici à un modèle de régression linéaire pénalisée pour lequel des connaissances a priori sont disponibles. Nous supposons que ces connaissances prennent la forme d'une certitude que certains régresseurs doivent appartenir au modèle. Dans ce cas particulier, leur pénalisation est mise à 0. La pénalisation des autres régresseurs potentiels reste inchangée et la même pour tous. Au lieu d'optimiser directement la fonction objectif du LASSO par un algorithme adapté, nous proposons tout d'abord de transformer le problème afin de diviser en deux parties la procédure d'optimisation. La première est une méthode des moindres carrés ordinaires, suivie d'une procédure LASSO classique en plus petite dimension. Ceci permet notamment d'améliorer l'estimation des paramètres ainsi que de réduire l'erreur de prédiction.
Dans la Section 2, nous présentons brièvement le LASSO et son extension. La Section 3 présente notre méthode appelée semi-LASSO et la Section 4 propose des simulations numériques afin de comparer les performances des méthodes 'glmnet' et semi-LASSO. Enfin, la Section 5 analyse les résultats obtenus.

# 2 LASSO et weighted LASSO

Supposons que nous avons un ensemble de $p$ variables explicatives $X_1, \ldots, X_p \in \mathbb{R}^p$ et une variable réponse $Y \in \mathbb{R}$ telles que :

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$ un bruit centré et $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p) \in \mathbb{R}^{p+1}$ les coefficients du modèle, potentiellement nuls. Supposons également avoir $n$ réalisations indépendantes de $X_1, \ldots, X_p$ et $Y$ : on note $\mathbf{x} = (x_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ la matrice des observations dont la première colonne est une colonne de 1, et $\mathbf{y} \in \mathbb{R}^n$ les observations de $Y$. Le problème de régression pénalisée LASSO s'écrit alors :

$$\min_{(\beta_0,\beta_1,...,\beta_p)} \quad \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}||_2^2 + \lambda \mathrm{P}(\boldsymbol{\beta}) \tag{1}$$

avec $\lambda$ positif le paramètre de régularisation et $\mathrm{P}(\boldsymbol{\beta})$ le terme de pénalisation LASSO défini par :

$$\mathrm{P}(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j| \tag{2}$$

Plus $\lambda$ est grand, plus le nombre de coefficients $\beta_j$ mis à 0 est élevé, créant ainsi un modèle parcimonieux. Pour résoudre le problème (1), une descente de gradient par coordonnées est généralement utilisée directement sur la fonction objectif à minimiser : c'est ainsi que fonctionne le package 'glmnet' de R. Dans l'Equation (2), tous les coefficients $\beta_j$ sont considérés de la même manière et pénalisés de façon égale. Une extension possible du LASSO consiste donc à ajouter un poids différent pour chaque coefficient dans le terme de pénalisation. L'objectif derrière cet ajout peut être double :

— On cherche à obtenir des propriété théorique sur l'estimateur $\hat{\boldsymbol{\beta}}$ : il s'agit de l'adaptive LASSO, proposé par *Zou* (2006). Dans ce cas, les poids sont liés aux données et modifiés de façon itérative dans le but d'obtenir certaines propriétés théoriques pouvant faire défaut à l'estimateur du LASSO original.
— On cherche à inclure des informations sur les régresseurs du modèle : il s'agit du weighted LASSO, utilisé par exemple par *Bergensen et al* (2011).

Dans ces cas-là, la fonction de pénalisation devient :

$$\mathrm{P}_{\boldsymbol{w}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_j |\beta_j| \tag{3}$$

avec $\boldsymbol{w} = (w_1, \ldots, w_p) \in \mathbb{R}^{+p}$ le vecteur des poids. Ce nouveau problème peut être résolu avec le même algorithme de descente de gradient que pour le LASSO classique, notamment en spécifiant ses propres poids avec le paramètre 'penalty.factor' dans la fonction 'glmnet'.

# 3 Semi-LASSO, un weighted LASSO pour inclure de l'information a priori

## 3.1 Ajouter des a priori dans le modèle

Les modèles de régression avec pénalisation sont très utilisés pour l'inférence de réseaux de gènes (*Sanguinetti et al* 2019), notamment grâce à leur propension à gérer de gros jeux de données où $n \ll p$, mais également car ils peuvent sélectionner quelques gènes intéressants en tant que régresseurs, opérant ainsi une sélection de variables parmi des milliers. Néanmoins, de plus en plus d'informations concernant les interactions entre les gènes sont aujourd'hui disponibles, par le biais d'expériences biologiques réalisées en laboratoire. Celles-ci peuvent être prises en compte dans le modèle grâce au weighted LASSO décrit dans la Section 2, avec des poids correctement choisis.
Nous nous focalisons ici sur un cas particulier de weighted LASSO, où les poids sont soit 0, c'est-à-dire pas de pénalisation, soit 1, comme dans le LASSO original. Plus concrètement, un poids de 0 signifie que la variable associée à ce poids sera automatiquement incluse dans le modèle et le coefficient $\beta_j$ associé sera génériquement non nul. Cela correspond donc aux variables pour lesquelles nous avons la certitude qu'elles sont liées à $Y$.
Les variables $\mathbf{X}$ sont ainsi divisées en 2 groupes :

- $G_K$ —K pour "Known"— représente l'ensemble des variables pour lesquelles nous sommes certains de leur appartenance au modèle. Le poids de chaque variable dans $G_K$ est mis à 0.
- $G_U$ —U pour "Unknown"— représente l'ensemble des variables pour lesquelles nous n'avons aucune information a priori sur leur appartenance au modèle. Leurs poids sont laissés à 1, chacune étant pénalisée de la même manière par le paramètre $\lambda$.

Il nous apparaît raisonnable de supposer que la majorité des modélisations faites à partir de données réelles comportent un terme constant ($\beta_0$). Ainsi nous décidons de placer la variable correspondant à ce paramètre dans le groupe $G_K$. Nous supposons également dans ce qui suit que le nombre de variables placées dans $G_K$ ne dépasse pas $n$ le nombre d'observations, sans quoi la méthode que nous proposons devient inapplicable.

## 3.2 Formulation mathématique

En utilisant les notations introduites précédemment, le problème que nous voulons résoudre se réécrit :

$$\min_{(\beta_0,\beta_1,\dots,\beta_p)} \quad \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}||_2^2 + \lambda \mathrm{P}_{\mathbf{1}}(\boldsymbol{\beta}) \tag{4}$$

avec

$$\begin{aligned} \mathrm{P}_{\mathbf{1}}(\boldsymbol{\beta}) &= \sum_{j=1}^{p} 1_{X_j \in G_U} |\beta_j| \\ &= \sum_{j,\, X_j \in G_U} |\beta_j| \end{aligned}$$

On note $\boldsymbol{\beta}_K$ (respectivement $\boldsymbol{\beta}_U$), le vecteur tel que $\boldsymbol{\beta}_K = (\beta_j,\, X_j \in G_K)$ (respectivement $\boldsymbol{\beta}_U = (\beta_j,\, X_j \in G_U)$). De la même manière, on note $\mathbf{x}_K$ (respectivement $\mathbf{x}_U$) la matrice contenant les $n$ observations des variables dans $G_K$ (respectivement $G_U$). L'équation (4) se réécrit alors :

$$\min_{\boldsymbol{\beta}_U} \big( \min_{\boldsymbol{\beta}_K} \big( \frac{1}{2n} ||\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}||_2^2 \big) + \lambda \sum_{j,\, X_j \in G_U} |\beta_j| \big) \tag{5}$$

$$= \min_{\boldsymbol{\beta}_U} \big( \frac{1}{2n} ||\mathbf{u} - \mathbf{v}\boldsymbol{\beta}_U||_2^2, + \lambda \sum_{j,\, X_j \in G_U} |\beta_j| \big) \tag{6}$$

avec $\mathbf{u}$ et $\mathbf{v}$ bien choisis. On reconnaît ici la fonction objectif du LASSO, dépendant uniquement de $\boldsymbol{\beta}_U$. La solution de l'équation (6) peut être trouvée numériquement, par exemple en utilisant la version classique du LASSO implémentée dans 'glmnet'. Une fois que $\hat{\boldsymbol{\beta}}_U$ est connu, nous pouvons en déduire $\hat{\boldsymbol{\beta}}_K$ en utilisant les valeurs obtenues pour $\hat{\boldsymbol{\beta}}_U$. Le pseudocode correspondant à cette procédure désignée dans la suite par semi-LASSO est donnée dans l'Algorithme 1.

Une solution alternative serait d'utiliser la fonction 'glmnet' directement en spécifiant des poids 0 et 1 dans l'argument 'penalty.factor'. Cependant, la méthode que nous proposons est plus efficace dans certaines situations pour les raisons suivantes.
En séparant les variables en deux groupes distincts avant d'optimiser numériquement la fonction objectif donnée Equation (4), nous avons deux problèmes de minimisation à résoudre. Le premier est un problème des moindres carrés pour une régression linéaire classique possédant une solution analytique bien connue, et le second est une régression avec pénalisation LASSO classique. Le deuxième problème est résolu numériquement, mais sur un espace de plus petite dimension, car il y a moins de variables, comparé à la méthode consistant à utiliser le weigthed LASSO de 'glmnet' directement. Nous verrons ultérieurement que le semi-LASSO permet d'obtenir une meilleure estimation des paramètres et de réduire l'erreur de prédiction.

# 4 Simulations numériques

## 4.1 Simulation des données

Afin de comparer notre méthode d'estimation de paramètres avec 'glmnet', nous effectuons des simulations numériques. Nous nous focalisons ici sur des simulations où $n = 100 \ll p = 500$. Nous avons également choisi de construire des variables $(X_1, \ldots, X_p)$ corrélées par

**Algorithm 1** Semi-LASSO

**Inputs** :
$G_K$ and $G_U$ the groups of variables
$\mathbf{x} \in \mathcal{M}_{n\times p}(\mathbb{R})$ the data matrix
$\mathbf{y} \in \mathbb{R}^n$ the vector of observed values for $Y$

**Output :**
$\boldsymbol{\beta}$ the vector of coefficients

**if** $G_K = \emptyset$ **then** ▷ No prior information is known
$\boldsymbol{\beta}$ obtained with classical LASSO

**else if** $G_U = \emptyset$ **then** ▷ All regressors are known
$\boldsymbol{\beta}$ obtained with classical regression

**else**
compute $\mathbf{x}_K$ and $\mathbf{x}_U$ from $\mathbf{x}$
add a column of 1 in $\mathbf{x}_K$ for the intercept estimation
compute $\mathbf{u}$ and $\mathbf{v}$
solve $\min_{\boldsymbol{\beta}_U} \left(\frac{1}{2n}||\mathbf{u} - \boldsymbol{\beta}_U\mathbf{v}||_2^2, +\lambda \sum_{j,\, X_j \in G_U} |\beta_j|\right)$ with glmnet to obtain $\hat{\boldsymbol{\beta}}_U$
deduce $\hat{\boldsymbol{\beta}}_K$
**end if**

blocs, en nous appuyant sur le travail de thèse de *Friguet* (2010). Enfin, seules $p_u = 40$ variables parmi les 500 sont réellement utilisées pour construire $Y$, ce qui signifie que seules 40 coordonnées de $\boldsymbol{\beta} \in \mathbb{R}^p$ sont non nulles.
50 jeux de données sont simulés, de façon à évaluer la variabilité provenant du bruit $\epsilon$ et de l'utilisation de la cross-validation pour trouver le paramètre de pénalisation optimal $\lambda$.

## 4.2 Introduire de l'a priori dans le modèle

Pour voir comment l'algorithme du semi-LASSO se comporte avec différents niveaux de connaissance a priori, nous ajoutons progressivement des variables du groupe $G_U$ au groupe $G_K$. Ainsi, nous testons 11 scenarii différents en commençant par le cas où $G_K = \emptyset$, aucun régresseur n'étant connu au préalable. Nous ajoutant ensuite 4 variables, correspondant à 10% des vrais régresseurs, dans $G_K$ plusieurs fois, pour atteindre finalement $G_K = (X_{j_1}, \ldots, X_{j_{p_u}})$. Pour chaque scenario, nous estimons $\boldsymbol{\beta}$, à la fois avec le semi-LASSO et le weighted LASSO de 'glmnet'.

## 4.3 Critères de comparaison utilisés

Afin de comparer les résultats obtenus avec le semi-LASSO et 'glmnet', nous utilisons les mesures suivantes :

— La sensibilité, définie par se $= \frac{TP}{TP+FN} = \frac{|\{\beta_j \neq 0\} \bigcap \{\widehat{\beta_j} \neq 0\}|}{|\{\beta_j \neq 0\}|}$, avec TP les vrais positifs, autrement dit le nombre de vrais régresseurs correctement sélectionnés par la méthode, et FN les faux négatifs, autrement dit le nombre de vrais régresseurs qui n'ont pas été retrouvé par la méthode. Ce taux mesure la proportion de régresseurs trouvés par la méthode parmi tous ceux étant réellement dans le modèle.

— La spécificité, définie par sp $= \frac{TN}{TN+FP} = \frac{|\{\beta_j = 0\} \bigcap \{\widehat{\beta_j} = 0\}|}{|\{\beta_j = 0\}|}$, avec TN les vrais négatifs, c'est-à-dire le nombre de régresseurs ne faisant pas partie du modèle et qui ne sont effectivement pas sélectionnés en tant que tels par la méthode, et FP les faux positifs, c'est-à-dire les régresseurs ne faisant pas partie du modèle mais malgré tout sélectionnés par la méthode. Ce taux mesure la proportion de variables ne faisant pas partie du modèle et qui sont correctement exclues du modèle par la méthode considérée.

D'autres critères permettant d'évaluer la performance de prédiction des deux méthodes seront présentées lors de l'exposé.

# 5 Résultats

La Figure 1 présente la sensibilité et la spécificité obtenues pour chacune des méthodes selon le niveau de connaissances lors de l'estimation des paramètres. Chaque boxplot correspond aux 50 estimations de paramètres, une pour chaque jeu de données. Nous pouvons tout d'abord remarquer que la sensibilité augmente avec la connaissance a priori, ce qui semble cohérent avec le fait que l'on force des variables pertinentes à être incluses dans le modèle. Lorsque le niveau de connaissance a priori ne dépasse pas 80%, le semi-LASSO semble plus efficace que 'glmnet' pour trouver les vrais régresseurs. Avec cette méthode, nous enlevons tout d'abord l'influence des variables de $G_K$ sur $Y$ avant d'utiliser un LASSO classique sur les variables restantes de $G_U$. Nous pouvons raisonnablement supposer que cette étape préalable permet au LASSO d'être ensuite plus efficace pour trouver les vrais régresseurs parmi les variables restantes.
Si nous nous intéressons à la spécificité, nous pouvons observer que celle-ci est proche de 1 pour les deux méthodes considérées. La méthode 'glmnet' donne de meilleurs résultats que le semi-LASSO, en particulier quand la connaissance a priori augmente. Encore une fois, cela est probablement lié à la première étape de l'algorithme du semi-LASSO : en enlevant l'influence des variables de $G_K$ lorsque la connaissance a priori est grande, nous forçons le LASSO à nous fournir des régresseurs parmi les variables restantes, alors même qu'il y en a de moins en moins à trouver. De plus en plus de variables non pertinentes sont donc rajoutées dans le modèle, faisant baisser la spécificité du semi-LASSO. Dans 'glmnet', l'optimisation de la fonction objectif est faite sur l'espace entier des $p$ variables, ce qui limite le nombre de faux positifs.

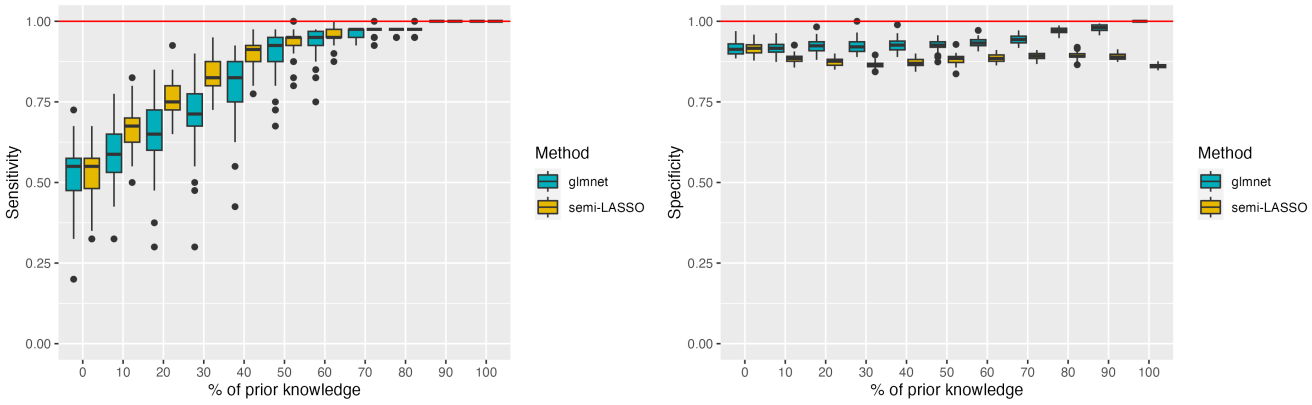



FIGURE 1 – Sensibilité et spécificité en fonction du niveau de connaissances lors de l'estimation des paramètres, pour les méthodes semi-LASSO et 'glmnet'.

# 6 Conclusion

Nous avons présenté une nouvelle méthode mettant en oeuvre un weighted LASSO, conçue spécifiquement pour intégrer des régresseurs connus au préalable dans le modèle. Il s'appuie sur le package R 'glmnet', mais n'utilise pas l'argument 'penalty.factor' permettant de spécifier les poids. Au lieu de cela, nous transformons d'abord le problème afin de diviser le problème d'optimisation en deux parties : une méthode des moindres carrés et un LASSO en plus petite dimension. Cela permet d'obtenir de meilleurs résultats notamment sur le nombre de vrais régresseurs détectés. Il faut malgré tout souligner que notre méthode ne peut s'appliquer que si le niveau d'a priori n'est pas trop élévé : $|G_K| \leq n$. Enfin, le semi-LASSO augmente le nombre de faux positifs par rapport à 'glmnet'.
Suivant le problème que l'utilisateur cherche à résoudre, il est laissé à sa discrétion d'utiliser le semi-LASSO ou 'glmnet' en ayant conscience des limites et avantages de chacune des deux méthodes.

# Bibliographie

Sanguinetti, G. , Huynh-Thu, V. *et al* (2019), Gene Regulatory Networks : Methods and Protocols, *Springer New York.*

Tibshirani, R. (1996),Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B : Statistical Methodology*

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418-1429.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software.*

Bergersen, L. C., Glad, I. K., Lyng, H. (2011). Weighted lasso with data integration. *Statistical*

*applications in genetics and molecular biology*, 10(1). Cassan, O., Lecellier, C. H., Martin, A., Brehelin, L., Lèbre, S. (2023). Optimizing data integration improves Gene Regulatory Network inference in Arabidopsis thaliana. bioRxiv, 2023-09.

Friguet, C. (2010). Impact de la dépendance dans les procédures de tests multiples en grande dimension (Doctoral dissertation, Agrocampus-Ecole nationale supérieure d'agronomie de rennes).

# Processus

# A spectral analysis approach for estimation of a Noisy Hawkes process

Anna Bonnet [1], Félix Cheysson [2], Miguel Martinez Herrera [3] & Maxime Sangnier[4]

[1] *Sorbonne Université, France, anna.bonnet@sorbonne-universite.fr*
[3] *Gustave Eiffel University, France, felix.cheysson@univ-eiffel.fr*
[3] *Sorbonne Université, France, miguel.martinez_herrera@sorbonne-universite.fr*
[4] *Sorbonne Université, France, maxime.sangnier@sorbonne-universite.fr*

**Résumé.** Les méthodes classiques d'estimation pour les processus de Hawkes auto-excitants sont fondées sur l'hypothèse que les événements observés correspondent à une réalisation d'un processus de Hawkes, sans perturbation aucune. Néanmoins, il est raisonnable d'envisager qu'en pratique les observations sont bruitées, en un sens dépendant du contexte. Il devient alors essentiel de modéliser la source de bruit et d'adapter les procédures d'inférence pour obtenir des estimations pertinentes d'un tel processus, dit bruité. Bien que plusieurs modélisations existent, nous considérons dans ce travail un mécanisme de bruit consistant en l'ajout d'observations qui proviennent d'un processus externe. Plus précisément, on suppose que les événements correspondent à la réunion de points provenant d'un processus de Hawkes et d'un processus de Poisson indépendants, sans que l'origine de ces points ne soit connue. Puisque dans ce contexte les méthodes canoniques d'estimation, comme le maximum de vraisemblance ou l'algorithme *Expectation-Maximisation*, sont soit impossibles à mettre en œuvre soit numériquement trop coûteuses, nous proposons une nouvelle procédure d'inférence fondée sur l'analyse spectrale des moments de second ordre du processus de Hawkes bruité. Nos contributions incluent une caractérisation du spectre de Bartlett grâce à sa densité spectrale, ainsi que des conditions d'identifiabilité du modèle statistique dans les cas de processus univariés et bivariés à interactions exponentielles. Un nouvel estimateur construit sur la maximisation de la log-vraisemblance spectrale est présenté et son comportement est analysé numériquement sur des données synthétiques. Il est ainsi montré qu'en plus d'être implémentable sans connaître l'origine des événements (processus de Hawkes ou de Poisson), l'estimateur proposé estime correctement les deux processus.

**Mots-clés.** Processus de Hawkes, processus ponctuel, analyse spectrale, estimation paramétrique, superposition, identifiabilité.

**Abstract.** Classical estimation methods for Hawkes processes rely on the assumption that observed event times are indeed a realisation of a Hawkes process, without considering any potential perturbation of the model. However, in practice, observations are often altered by some noise, the form of which depends on the context. It is then required to model the alteration mechanism in order to infer accurately such a noisy Hawkes process. While several models exist, we consider, in this work, the observations to be the indistinguishable union of event times coming from a Hawkes process and from an independent Poisson process. Since standard inference methods (such as maximum likelihood or Expectation-Maximisation) are either unworkable or numerically prohibitive in this context, we propose an estimation procedure based on the spectral analysis of second order properties of the noisy Hawkes process.

Novel results include a characterisation of the Bartlett spectrum thanks to its spectral density, and sufficient conditions for the identifiability of the ensuing statistical model with exponential interaction functions for both univariate and bivariate processes. A new estimator based on maximising the spectral log-likelihood is then described, and its behaviour is numerically illustrated on synthetic data. Besides being free from knowing the source of each observed time (Hawkes or Poisson process), the proposed estimator is shown to perform accurately in estimating both processes.

**Keywords.** Hawkes process, point process, spectral analysis, parametric estimation, superposition, identifiability.

# 1 Introduction

Hawkes processes, introduced in Hawkes (1971), are a class of point processes that have been originally used to model self-exciting phenomena and more recently other types of past-dependent behaviours. Their fields of applications are wide and include for instance seismology (Ogata, 1988, 1998), neuroscience (Chornoboy et al., 1988; Lambert et al., 2018), criminology (Olinde and Short, 2020), finance (Embrechts et al., 2011; Bacry et al., 2015) and biology (Gupta et al., 2018), to mention a few. Consequently, there has been a deep focus on estimation techniques for Hawkes processes. Among them, let us mention maximum likelihood approaches (Ogata, 1978; Ozaki, 1979; Guo et al., 2018), methods of moments (Da Fonseca and Zaatour, 2013), least-squares contrast minimisation (Reynaud-Bouret et al., 2014; Bacry et al., 2020), Expectation-Maximisation (EM) procedures (Lewis and Mohler, 2011), methods using approximations through autoregressive models (Kirchner, 2017).

As a consequence of this, there has been a deep focus on estimation techniques for Hawkes processes through maximum likelihood estimation (Ogata, 1978; Ozaki, 1979; Guo et al., 2018), method of moments (Da Fonseca and Zaatour, 2013), least-squares contrast minimisation (Bacry et al., 2020), EM procedure (Lewis and Mohler, 2011), using approximations through autoregressive models (Kirchner, 2017), by decomposition on histogram basis (Reynaud-Bouret et al., 2014).

All of these methods assume that the history of the point process has been accurately observed, which is untrue in many applications. For example, when reading spike train for the study of neuronal activation networks, it is likely to detect additional points not corresponding to real events.

Although the study of noised observations is a common issue in other contexts, it has been scarcely studied in the point processes setting. The works of Lund and Rudemo (2000) focus on a general model where a point process' event times are randomly thinned, displaced and an additional external "noise" is added in the form of an exogeneous point process.

In a recent contribution by Cheysson and Lang (2022), the authors employ the spectral analysis of point processes to study a noised version of a Hawkes process. The model consists on an aggregate count of the real event times within specific time intervals instead of precise

occurrences in time. By leveraging the properties of the Bartlett spectrum, they propose an estimator obtained through means of maximisation of the spectral log-likehood. They illustrate its efficacy on synthetic and real-world data on a parametric context for different kernel functions.

In this communication, we consider the undistinguishable superposition of a Hawkes process and a homogeneous Poisson process. We derive a parametric estimation procedure by leveraging the general expression for the spectral density of this model. This approach has the advantage of not needing any knowledge on the source of the observed event times. By working with the exponential kernel function for the Hawkes process, we provide sufficient conditions for the identifiability of our model. Our numerical procedure is implemented in Python and we carry out a numerical study on synthetic data for all different identifiable scenarios illustrating the accuracy of our estimations.

## 2 Mathematical setting

Let $H = (H_1, \dots, H_d)$ be a stationary multivariate Hawkes process on $\mathbb{R}$ defined by its conditional intensity functions $\lambda_i^H$ ($i \in \{1, \dots, d\}$): for all $t \in \mathbb{R}$,

$$\lambda_i^H(t) = \mu_i + \sum_{j=1}^{d} \int_{-\infty}^{t} h_{ij}(t-s)\, H_j(\mathrm{d}s) = \mu_i + \sum_{j=1}^{d} \sum_{T_k^{H_j} \leq t} h_{ij}(t - T_k^{H_j})\,, \tag{1}$$

where $\mu_i > 0$ is the baseline intensity of process $H_i$, $h_{ij} > 0$ is the interaction or kernel function describing the effect of process $H_j$ on process $H_i$, and $(T_k^{H_j})_{k\geq 1}$ denotes the event times of $H_j$.

By defining the matrix $S^+ = (\|h_{ij}\|_1)_{ij}$ where

$$\|h_{ij}\|_1 = \int_{-\infty}^{+\infty} h_{ij}(t)\,\mathrm{d}t\,,$$

the stationarity condition of $H$ reduces to controlling the spectral radius of $S^+$: $\rho(S^+) < 1$ (Brémaud and Massoulié, 1996). In this communication, we consider the exponential kernel function defined for any integers $i, j$ as:

$$h_{ij}(t) = \alpha_{ij}\beta_i \mathrm{e}^{-\beta_i t}\,, \quad \text{for } t > 0, \tag{2}$$

with $\alpha_{ij} \geq 0$ and $\beta_i > 0$, and so $S^+ = (\alpha_{ij})_{ij}$.

Our goal is to study a noisy version of the Hawkes process where the sequence of event times $(T_k^H)_{k\geq 1}$ of $H$ is contaminated by the event times from another process, which is chosen to be a multivariate Poisson process.

Formally, let $P = (P_1, \dots, P_d)$ be a multivariate process made up with $d$ independent homogeneous Poisson processes on $\mathbb{R}$ (denoted $P_1, \dots, P_d$) with shared intensity $\lambda_0 > 0$, supposed to be independent from $H$. We note the event times $(T_k^{P_i})_{k\geq 1}$.

We consider then the point process $N = (N_1, \dots, N_d)$ defined as the superposition of $H$ and $P$. The sequence of event times $(T_k^{N_i})_{k\geq 1}$ of $N_i$ is the ordered union of $(T_k^{H_i})_{k\geq 1}$ and $(T_k^{P_i})_{k\geq 1}$. Throughout this paper we will refer to $N$ as the **noisy Hawkes process**.

Our goal is to estimate both processes parameters (i.e. the baselines $\mu_i$, the kernels $h_{ij}$ and the shared Poisson intensity $\lambda_0$) from the sole observation of $(T_k^{N_i})_{k\geq 1}$. Our method consists in characterising a multivariate point process by its matrix-valued spectral density function, denoted $\mathbf{f} : \mathbb{R} \to \mathbb{C}^{d\times d}$ (with for all $\omega \in \mathbb{R}$, $\mathbf{f}(\omega) = (f_{ij}(\omega))_{1\leq i,j\leq d}$), which is related to second-order measures (Bartlett, 1963).

We define the second order moment measures $M_{ij}^N$ for any $i, j$ as:

$$M_{ij}^N(A, B) \coloneqq \mathbb{E}[N_i(A)N_j(B)] = \int_{A\times B} N_i(\mathrm{d}x)\, N_j(\mathrm{d}y)\,.$$

As the process is stationary, $M_{ij}^N$ can be decomposed in a product of $\ell_{\mathbb{R}}$ and a so-called reduced measure $\breve{M}_{ij}^N$, such that for any bounded measurable function $g$ of bounded support

$$\int_{\mathbb{R}^2} g(x, y)\, M_{ij}^N(\mathrm{d}x, \mathrm{d}y) = \int_{\mathbb{R}}\int_{\mathbb{R}} g(x, x + u)\, \ell_{\mathbb{R}}(\mathrm{d}x)\breve{M}_{ij}^N(\mathrm{d}u)\,. \tag{3}$$

The spectral density function can then be expressed as:

$$f_{ij}^N(\omega) = \int_{\mathbb{R}} \mathrm{e}^{-2\pi i x\omega}\, \breve{M}_{ij}^N(\mathrm{d}x) - \mathbb{E}[\lambda_i^H(0)]\mathbb{E}[\lambda_j^H(0)]\mathbb{1}_{\omega=0}, \quad \forall\omega \in \mathbb{R}\,. \tag{4}$$

Given some observed multivariate times $(T_k^N)_{k\geq 1}$ (in a prescribed time window $[0, T]$), for any pair $(i, j) \in \{1, \dots, d\}^2$, spectral density functions can be approximated by the cross-periodograms, defined for all $\omega \in \mathbb{R}$ by:

$$I_{ij}^T(\omega) = \sum_{k=1}^{N_i(T)} \sum_{l=1}^{N_j(T)} \mathrm{e}^{-2\pi i\omega(T_k^{N_i} - T_l^{N_j})}\,, \tag{5}$$

where $N_i(t) = N_i([0, t))$. The matrix-valued function $\mathbf{I}^T : \omega \in \mathbb{R} \mapsto (I_{ij}^T(\omega))_{1\leq i,j\leq d}$ can be computed regardless of the knowledge of the source of the event times.

In the scope of statistical inference, a parametric model for spectral density functions is considered:

$$\mathcal{P} = \left\{\mathbf{f}_\theta^N : \mathbb{R} \to \mathbb{C}^{d\times d}, \theta = (\mu, h, \lambda_0) \in \Theta\right\}\,.$$

Then, the so-called Whittle estimator $\mathbf{f}_{\hat{\theta}}^N$ of $\mathbf{f}$ (Whittle, 1952), can be obtained by maximising the approximate spectral log-likelihood (Brillinger, 2012; Düker and Pipiras, 2019; Villani et al., 2022):

$$\ell_T(\theta) = -\frac{1}{T}\sum_{k=1}^{M} \left(\log(\det(\mathbf{f}_\theta^N(\omega_k))) + \mathrm{Tr}\left(\mathbf{f}_\theta^N(\omega_k)^{-1}\mathbf{I}^T(\omega_k)\right)\right), \tag{6}$$

where det and Tr are respectively the determinant and trace of matrices, $\omega_k = k/T$, $k \in \{1, 2, \dots\}$ and $M$ is a hyperparameter.

# 3 Spectral density and estimation

## 3.1 Superposition of processes

As our model focuses on the superposition of two point processes, the following general result can be first established.

**Proposition 3.1.** *Let $X, Y$ be two independent stationary multivariate point processes on $\mathbb{R}$, admitting respective spectral density matrices $\mathbf{f}^X$, $\mathbf{f}^Y$.*

*Then $N = X + Y$ also admits a spectral density matrix and:*

$$\mathbf{f}^N = \mathbf{f}^X + \mathbf{f}^Y \,. \tag{7}$$

We can then explicit the spectral matrix $\mathbf{f}^N$ by leveraging Proposition 3.1. Indeed, under stationarity conditions, the spectral matrix $\mathbf{f}^H$ of a Hawkes process is known to depend on the Fourier transform of the interaction functions and on the mean densities (Daley and Vere-Jones, 2003, Example 8.3(c)). More precisely, we define the matrix of Fourier transform interactions as:

$$\tilde{h}(\omega) := (\tilde{h}_{ij}(\omega))_{ij} \,,$$

where $\tilde{h}_{ij}$ denotes the Fourier transform of $h_{ij}$.

The mean intensities $m_i^H$ of each process $H_i$ can be obtained by:

$$\begin{pmatrix} m_1^H \\ \vdots \\ m_d^H \end{pmatrix} = (I_d - \tilde{h}(0))^{-1} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} \,,$$

where $I_d$ is the identity matrix of dimension $d$, and the spectral matrix by (Daley and Vere-Jones, 2003, Equation 8.3.11):

$$\mathbf{f}^H(\omega) = (I_d - \tilde{h}(-\omega))^{-1} \operatorname{Diag}(m^H)(I_d - \tilde{h}(\omega)^T)^{-1} \,, \tag{8}$$

with $\operatorname{Diag}(m^H)$ defined as the diagonal matrix formed by the mean intensities $m_i^H$.

As this expression is still true for a homogeneous Poisson process (with $\tilde{h} \equiv 0$), the spectral matrix of a noisy Hawkes process $N$ reads:

$$\mathbf{f}(\omega) = \mathbf{f}^H(\omega) + \lambda_0 I_d \,, \tag{9}$$

with $\mathbf{f}^H$ defined by Equation (8). With this result, we can establish an estimation method by optimising the spectral log-likelihood in Equation (6).

## 3.2 Identifiability in the univariate framework

When working in the univariate framework, we define the exponential model for noisy Hawkes process as:

$$\mathcal{Q} = \left\{ f_\theta^N : \theta := (\mu, \alpha, \beta, \lambda_0) \in \mathbb{R}_{>0} \times [0,1) \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \right\} ,$$

where $f_\theta^N$ is the single spectral density function of the candidate noisy Hawkes process, $\alpha$ and $\beta$ are the parameters of the single kernel (see Equations (2) and (4)).

By leveraging Proposition 3.1, the spectral density reduces to:

$$f_\theta^N(\omega) = \frac{\mu}{1-\alpha}\alpha\beta^2(2-\alpha)\left(\frac{1}{(\beta(1-\alpha))^2 + (2\pi\omega)^2}\right) + \left(\frac{\mu}{1-\alpha} + \lambda_0\right) , \tag{10}$$

and we obtain the following result concerning identifiability.

**Proposition 3.2.** *The model $\mathcal{Q}$ is not identifiable. In particular, for any admissible parameter $\theta = (\mu, \alpha, \beta, \lambda_0)$ there exists an infinite number of parameters $\theta'$ such that $f_\theta = f_{\theta'}$.*

*However, if we assume that one of the four quantities in $\theta$ is known, then the reduced model defined by the remaining triplet of parameters is identifiable.*

## 3.3 Identifiability in the bivariate framework

In the bivariate framework, let now consider the model:

$$\mathcal{Q}_\Lambda = \left\{ \mathbf{f}_\theta : \theta := (\mu, \alpha, \beta, \lambda_0) \in \mathbb{R}_{>0}^2 \times \Lambda \times \mathbb{R}_{>0}^2 \times \mathbb{R}_{>0}, \rho(\alpha) < 1 \right\} ,$$

where $\Lambda \subset [0,1)^{2\times 2}$. Since identifiability when $\Lambda = [0,1)^{2\times 2}$ deserves an intricate analysis, only two particular situations are presented below.

Proposition 3.3 presents two non-identifiable scenarios and in Proposition 3.4 we establish sufficient conditions for identifiability of $\mathcal{Q}_\Lambda$.

**Proposition 3.3.** *The model $\mathcal{Q}_\Lambda$ is not identifiable in both situations:*

1. $\Lambda = \{\alpha \in [0,1)^{2\times 2} : \alpha_{12} = \alpha_{21} = 0\}$, *that is for diagonal matrices $\alpha$ (with possibly null entries).*
2. $\Lambda = \{\alpha \in [0,1)^{2\times 2} : \alpha_{11}\alpha_{12} = 0, \alpha_{21}\alpha_{22} > 0$ *or* $\alpha_{11}\alpha_{12} > 0, \alpha_{21}\alpha_{22} = 0\}$, *that is for matrices $\alpha$ with a null row and a row with positive entries.*

**Proposition 3.4.** *The model $\mathcal{Q}_\Lambda$ is identifiable in both situations:*

1. $\Lambda = \{\alpha \in [0,1)^{2\times 2} : \alpha_{11} \geq 0, \alpha_{21} > 0, \alpha_{12} = \alpha_{22} = 0$ *or* $\alpha_{11} = \alpha_{21} = 0, \alpha_{12} \geq 0, \alpha_{22} > 0\}$, *that is for matrices $\alpha$ with a null column and a positive entry on the antidiagonal.*
2. $\Lambda = \{\alpha \in [0,1)^{2\times 2} : \alpha_{11} > 0, \alpha_{22} > 0, \alpha_{12}\alpha_{21} = 0\}$, *that is for matrices $\alpha$ with a positive diagonal and at least a null entry on the antidiagonal.*

# 4 Numerical results

This section depicts a simple numerical illustration of the behaviour of the proposed estimator $\hat{\theta}$, obtained by maximising the spectral log-likelihood $\ell_T$ (see Equation (6)) thanks to the L-BFGS-B method implemented in the `scipy` Python package. A univariate process is considered, along with the exponential model $\mathcal{Q}$. It results that $\hat{\theta} = (\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\lambda}_0)$, that is the base intensity $\hat{\mu}$ of the Hawkes process, the interaction magnitude $\hat{\alpha}$ and the decay rate $\hat{\beta}$ of the kernel; and the intensity $\hat{\lambda}_0$ of the noise.

Observations are simulated according to an exponential Hawkes process with $(\mu, \alpha, \beta) = (1.0, 0.5, 1.0)$ and a Poisson process with $\lambda_0 = 1.2$ (so the target parameter is $\theta^\star = (\mu, \alpha, \beta, \lambda_0) = (1.0, 0.5, 1.0, 1.2)$) and the behaviour of $\hat{\theta}$ is analysed via the relative error $\frac{\|\hat{\theta} - \theta^\star\|_2}{\|\theta^\star\|_2}$ averaged over 50 repetitions. This setting is without loss of generality and, in particular all conclusions are consistent with other settings such as $\lambda_0 = 0.4$ and $\lambda_0 = 2.0$ (either decreased or increased noises).

Since the considered model $\mathcal{Q}$ is not identifiable (see Proposition 3.2) unless fixing a parameter, four situations are presented (see Figure 1): estimating i) $(\alpha, \beta, \lambda)$ with $\mu$ fixed, ii) $(\mu, \beta, \lambda)$ with $\alpha$ fixed, iii) $(\mu, \alpha, \lambda)$ with $\beta$ fixed, iv) $(\mu, \alpha, \beta)$ with $\lambda_0$ fixed. The fixed parameter is set to the target value each time.

Figure 1 presents the trend of the relative errors with respect to the number of time events (top panels) and to the computation time (bottom panels). As expected, they decrease to 0, suggesting that the estimator $\hat{\theta}$ converges (in quadratic mean) to $\theta^\star$ as the horizon $T$ goes to infinity.

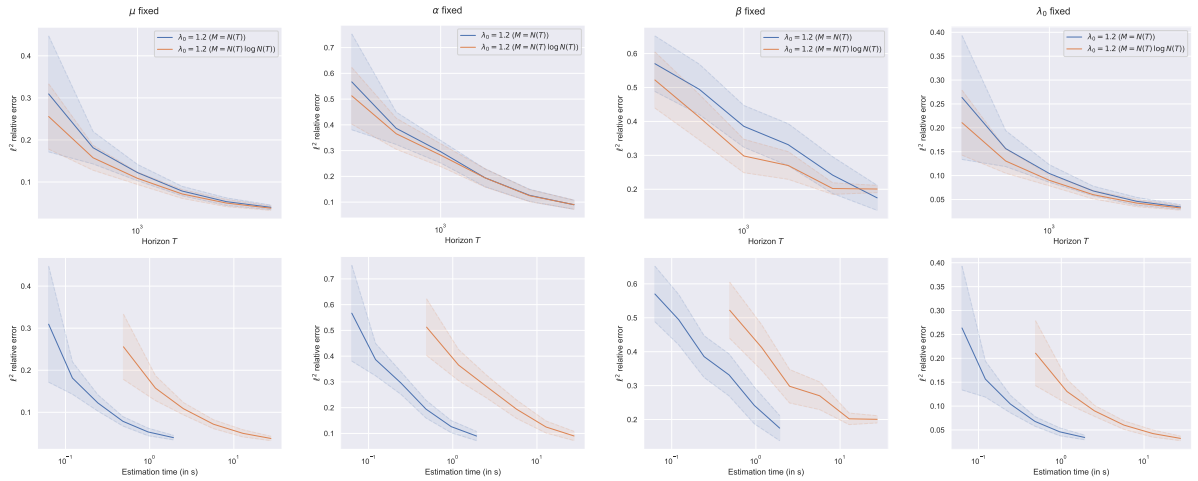



Figure 1: Relative estimation error respectively for $\mu$, $\alpha$, $\beta$ and $\lambda_0$ fixed (columns from left to right) with respect to the time horizon $T$ (top) and the computation time (bottom).

Moreover, as suggested in (Pham, 1996), the choice of the hyperparameter $M$ is important. We follow (Cheysson and Lang, 2022), that proposes to consider $M$ increasing with the number of observed times $N(T)$. Figure 1 illustrates the trend of the relative errors when $M = N(T)$ (blue curves) and when $M = N(T) \log N(T)$ (orange curves). It appears, first,

that the relative errors are comparable for a given number of points $N(T)$ (top panels); second, that the case $M = N(T) \log N(T)$ requires a fitted time which is roughly 10 times greater than that for $M = N(T)$. Overall, it seams preferable to choose $M = N(T)$.

# 5 Conclusion

In this communication, we proposed an estimation procedure for the superposition of a Hawkes process and a Poisson process when missing information about the source of each process' event times. Our first contribution is providing explicit expressions for the spectral density of any superposition of two independent point processes. We established sufficient conditions for the identifiability of such a model. We then developed and implemented an estimator through the maximisation of the spectral log-likelihood in such scenarios and illustrated the efficiency of our method on synthetic data.

Future work include a detailed study of the multivariate setting, both from the theoretical and from the numerical points of view. On the first hand, it will be tried to extend the identifiability results to dimension $d \geq 3$. On the other hand, our estimation method will be numerically assessed on real datasets regarding neuronal activity with multiple subprocesses.

# References


Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J. (2020), "Sparse and low-rank multivariate Hawkes processes," *Journal of Machine Learning Research*, 21, 1–32.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015), "Hawkes Processes in Finance," *Market Microstructure and Liquidity*, 01, 1550005.

Bartlett, M. S. (1963), "The Spectral Analysis of Point Processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, 25, 264–296.

Brillinger, D. R. (2012), *Statistical Inference for Stationary Point Processes*, New York, NY: Springer New York, 499–543.

Brémaud, P. and Massoulié, L. (1996), "Stability of nonlinear Hawkes processes," *The Annals of Probability*, 24, 1563–1588.

Cheysson, F. and Lang, G. (2022), "Spectral estimation of Hawkes processes from count data," *The Annals of Statistics*, 50, 1722 – 1746.

Chornoboy, E. S., Schramm, L. P., and Karr, A. F. (1988), "Maximum likelihood identification of neural point process systems," *Biological Cybernetics*, 59, 265–275.

Da Fonseca, J. and Zaatour, R. (2013), "Hawkes Process: Fast Calibration, Application to Trade Clustering, and Diffusive Limit," *Journal of Futures Markets*, 34, 548–579.

Daley, D. and Vere-Jones, D. (2003), *An introduction to the theory of point processes. Vol. I*, New York: Springer-Verlag, second edition.

Düker, M.-C. and Pipiras, V. (2019), "Asymptotic results for multivariate local Whittle estimation with applications," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*.

Embrechts, P., Liniger, T., and Lin, L. (2011), "Multivariate Hawkes processes: an application to financial data," *Journal of Applied Probability*, 48, 367–378.

Guo, X., Hu, A., Xu, R., and Zhang, J. (2018), "Consistency and Computation of Regularized MLEs for Multivariate Hawkes Processes," Preprint at https://arxiv.org/abs/1810.02955.

Gupta, A., Farajtabar, M., Dilkina, B., and Zha, H. (2018), "Discrete Interventions in Hawkes Processes with Applications in Invasive Species Management," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, International Joint Conferences on Artificial Intelligence Organization.

Hawkes, A. (1971), "Spectra of Some Self-Exciting and Mutually Exciting Point Processes," *Biometrika*, 58, 83–90.

Kirchner, M. (2017), "An estimation procedure for the Hawkes process," *Quantitative Finance*, 17, 571–595.

Lambert, R., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N., and Reynaud-Bouret, P. (2018), "Reconstructing the functional connectivity of multiple spike trains using Hawkes models," *Journal of Neuroscience Methods*, 297, 9–21.

Lewis, E. and Mohler, G. (2011), "A Nonparametric EM Algorithm for Multiscale Hawkes Processes," *Journal of Nonparametric Statistics*, 1, 1–20.

Lund, J. and Rudemo, M. (2000), "Models for Point Processes Observed with Noise," *Biometrika*, 87, 235–249.

Ogata, Y. (1978), "The asymptotic behaviour of maximum likelihood estimators for stationary point processes," *Annals of the Institute of Statistical Mathematics*, 30, 243–261.

— (1988), "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes," *Journal of the American Statistical Association*, 83, 9–27.

— (1998), "Space-Time Point-Process Models for Earthquake Occurrences," *Annals of the Institute of Statistical Mathematics*, 50, 379–402.

Olinde, J. and Short, M. (2020), "A Self-limiting Hawkes Process: Interpretation, Estimation, and Use in Crime Modeling," in *2020 IEEE International Conference on Big Data (Big Data)*.

Ozaki, T. (1979), "Maximum likelihood estimation of Hawkes' self-exciting point processes," *Annals of the Institute of Statistical Mathematics*, 31, 145–155.

Pham, D. (1996), "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Transactions on Signal Processing*, 44, 2768–2779.

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., and Tuleau-Malot, C. (2014), "Goodness-of-Fit Tests and Nonparametric Adaptive Estimation for Spike Train Analysis," *The Journal of Mathematical Neuroscience*, 4, 3.

Villani, M., Quiroz, M., Kohn, R., and Salomone, R. (2022), "Spectral Subsampling MCMC for Stationary Multivariate Time Series with Applications to Vector ARTFIMA Processes," *Econometrics and Statistics*.

Whittle, P. (1952), "Some results in time series analysis," *Scandinavian Actuarial Journal*, 1952, 48–60.

# Estimation of subcritical Galton Watson processes with correlated immigration

Yacouba Boubacar Maïnassara [1] & Landy Rabehasaina [2]

[1] *Laboratoire de Matériaux Céramiques et de Mathématiques, Université Polytechnique Hauts-de-France, INSA Hauts-de-France, Valenciennes, France, Email: Yacouba.BoubacarMainassara@uphf.fr*

[2] *Laboratoire de Mathématiques de Besançon, Université de Franche Comté, Besançon, France, Email: lrabehas@univ-fcomte.fr*

**Résumé.** Nous considérons un processus observé de Galton Watson $\{Y_n,\ n \in \mathbb{Z}\}$ avec immigration modélisé par un processus corrélé $\{\epsilon_n,\ n \in \mathbb{Z}\}$. Nous présentons des résultats d'estimation du taux de reproduction et l'espérance de l'immigration dans deux situations. La première est lorsque $\mathrm{Cov}(\epsilon_0, \epsilon_k) = 0$ pour $k$ supérieur à un certain $k_0$: nous fournissons un estimateur et prouvons un résultat de normalité asymptotique. Dans un deuxième temps, nous considérons le cas où $\{\epsilon_n,\ n \in \mathbb{Z}\}$ a une structure de corrélation générale. Sous des hypothèses de mélange, nous déterminons un estimateur pour le taux de reproduction et nous montrons sa convergence en moyenne quadratique avec vitesse explicite. Lorsque le coefficient de mélange décroit suffisamment vite, un développement d'ordre 2 pour cet estimateur est établi. **Mots-clés.** Processus de Galton Watson, immigration, processus INAR.

**Abstract.** We consider an observed subcritical Galton Watson process $\{Y_n,\ n \in \mathbb{Z}\}$ with correlated stationary immigration process $\{\epsilon_n,\ n \in \mathbb{Z}\}$. Two situations are presented. The first one is when $\mathrm{Cov}(\epsilon_0, \epsilon_k) = 0$ for $k$ larger than some $k_0$: a consistent estimator for the reproduction and mean immigration rates is given, and a CLT is proved. The second one is when $\{\epsilon_n,\ n \in \mathbb{Z}\}$ has general correlation structure: under mixing assumptions, we exhibit an estimator for the logarithm of the reproduction rate and we prove that it converges in quadratic mean with explicit speed. In addition, when the mixing coefficients decrease fast enough, we provide and prove a two terms expansion for the estimator.

**Keywords.** Galton Watson processes, immigration, INAR processes.

# 1 Introduction and model

Estimation of parameters in a Galton Watson process with immigration has a long history: we refer to the seminal paper Klimko and Nelson (1978) for laying the ground and expliciting conditional least square estimators for the expectation of the reproduction and immigration sequences in the subcritical case. A central limit theorem for these estimators was later proved in Venkataraman (1982), using a time series point of view. Note that the link between such processes and the so called integer valued times series $INAR(1)$ processes has been exploited, see e.g. Al-Osh and Alzaid (1987) which studied such a process with particular distributions

for the reproduction and immigration sequences. A certain number of extensions for the model were later devised and studied. Wei and Winnicki (1990) considered the general critical and supercritical case and proved central limit theorems for (modified) weighted least square estimators. In Barczy et al (2021), the specific case where the immigration sequence has a regular variation distribution is considered, leading to asymptotic normality of the reproduction mean when the immigration mean is known. Generalization to two types processes have been recently investigated in Ispány et al (2014), Körmendi and Pap (2018), including estimations for the criticality parameter. Note that these references have one of the following constraint on the immigration process: first, some of them assume that its expectation is known (and appears in the expression of the estimator for the reproduction sequence); second, this immigration process is assumed to be a sequence of *independent* random variables, in addition to be identically distributed. Hence, we aim in this paper at considering a process which is stationary, but where some form of dependence is given. This particular feature appears not to have been studied in the literature.

We consider the following Galton Watson process with immigration as the stationary sequence $\{Y_n,\ n \in \mathbb{Z}\}$ satisfying

$$Y_{n+1} = \sum_{k=1}^{Y_n} \xi_{n+1,k} + \epsilon_{n+1}, \quad n \in \mathbb{Z}, \tag{1}$$

for some sequences $\{\xi_{n,k},\ n \in \mathbb{Z},\ k \in \mathbb{N}\}$ and $\{\epsilon_n,\ n \in \mathbb{Z}\}$, named thereafter the *reproduction* sequence and the *immigration* sequence, which are such that

- $\{\xi_{n,k},\ n \in \mathbb{Z},\ k \in \mathbb{N}\}$ and $\{\epsilon_n,\ n \in \mathbb{Z}\}$ are independent sequences,
- the reproduction sequence $\{\xi_{n,k},\ n \in \mathbb{Z},\ k \in \mathbb{N}\}$ is an i.i.d. (doubly indexed) sequence, with distribution of a generic r.v. denoted by $\xi$,
- the immigration process $\{\epsilon_n,\ n \in \mathbb{Z}\}$ is stationary and ergodic, with distribution that of a generic r.v. denoted by $\epsilon$,
- the moments $\lambda_0 := \mathbb{E}(\xi)$ and $m_0 := \mathbb{E}(\epsilon)$ of $\{\xi_{n,k},\ n \in \mathbb{Z},\ k \in \mathbb{N}\}$ and $\{\epsilon_n,\ n \in \mathbb{Z}\}$ are unknown.

The aim of the paper is to estimate the unknown reproduction rate $\lambda_0$ as well as the mean immigration $m_0$ from the observed sequence $\{Y_n,\ n \in \mathbb{Z}\}$. Furthermore, in order for the existence of the stationary process $\{Y_n,\ n \in \mathbb{Z}\}$ to exist, it is required that the subcritical case holds, namely that

$$\lambda_0 < 1. \tag{2}$$

To control the serial dependence of the stationary process $\{\epsilon_n,\ n \in \mathbb{Z}\}$, we introduce the strong mixing coefficients $\alpha_\epsilon(h)$ defined by

$$\alpha_\epsilon(h) = \sup_{A \in \mathcal{F}_{-\infty}^n, B \in \mathcal{F}_{n+h}^\infty} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|,$$

where $\mathcal{F}_{-\infty}^n = \sigma(\epsilon_u, u \leq n)$ and $\mathcal{F}_{n+h}^\infty = \sigma(\epsilon_u, u \geq n+h)$. Note that $\alpha_\epsilon(h)$ does not depend on $n \in \mathbb{Z}$ thanks to the stationarity of $\{\epsilon_n,\ n \in \mathbb{Z}\}$.

To establish the asymptotic properties of the proposed estimator of $\lambda_0$, the following assumptions are required.

**(A1):** The mixing coefficient $\alpha_\epsilon(\cdot)$ of $\{\epsilon_n,\ n \in \mathbb{Z}\}$ verifies the following summability condition:

$$\sum_{h=0}^{\infty} \lambda_0^{-h} \alpha_\epsilon(h)^{1-2/\beta} < \infty \quad \text{for some} \quad \beta > 2.$$

We next make an integrability assumption on the moments and covariances of the immigration process $\{\epsilon_n,\ n \in \mathbb{Z}\}$ and reproduction sequence $\{\xi_{n,k},\ n \in \mathbb{Z},\ k \in \mathbb{N}\}$. We use $\|\cdot\|$ to denote the Euclidean norm of a vector and for any (potentially matrix valued) random variable $X$, we will set $\|Y\|_p^p := \mathbb{E}\|X\|^p$ its $\mathbb{L}^p$ norm, with $p \geq 1$.

**(A2):** The following moment conditions hold:

$$\|\epsilon\|_{2\beta} = \left[\mathbb{E}\left|\epsilon\right|^{2\beta}\right]^{1/(2\beta)} < \infty \quad \text{and} \quad \|\xi\|_{2\beta} = \left[\mathbb{E}\left|\xi\right|^{2\beta}\right]^{1/(2\beta)} < \infty.$$

**(A3):** The covariance of the immigration process $\nu_h := \mathrm{Cov}(\epsilon_0, \epsilon_h)$ verifies:

$$\sum_{h=0}^{\infty} h\lambda_0^{-h} \left|\nu_{h+1}\right| < \infty.$$

Let us then define for all $n \in \mathbb{N}$ and $k \geq 0$

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \bar{Y}_{k+1,n} := \frac{1}{n}\sum_{i=1}^{n} Y_i Y_{i+k+1}.$$

# 2 Ultimately uncorrelated immigration

We suppose in this section that the immigration is no more correlated from an instant $k_0$, i.e. that the following assumption holds

**$(\mathbf{A3})_\mathbf{1}$:** There exists a known $k_0 \in \mathbb{N} \setminus \{0\}$ such that $\nu_k = 0$ for all $k \geq k_0$.

The main results are given as follows.

**Proposition 1** (Consistency)**.** Let us suppose that (**A1**), (**A2**), (**A3**) and $(\mathbf{A3})_\mathbf{1}$ hold. Then the following estimators

$$\hat{R}_{k_0,n} := \frac{\left[\bar{Y}_n\right]^2 - \bar{Y}_{k_0,n}}{\left[\bar{Y}_n\right]^2 - \bar{Y}_{k_0-1,n}}, \quad \hat{M}_{k_0,n} := \bar{Y}_n(1 - \hat{R}_{k_0,n}) \tag{3}$$

converge a.s. towards the unknown parameters $\lambda_0$ and $m_0$ as $n \to \infty$.

**Theorem 2** (Asymptotic normality)**.** Under the previous assumptions, the following CLT holds:

$$\sqrt{n}[\hat{R}_{k_0,n} - \lambda_0, \hat{M}_{k_0,n} - m_0]' \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \Omega_{k_0}), \tag{4}$$

where $\Omega_{k_0}$ is explicit.

# 3 General correlated immigration

In this case, additional Assumptions are required.

$(\mathbf{A3})_\mathbf{2}$: The mixing coefficient $\alpha_\epsilon(\cdot)$ has exponential decrease.

(**A4**): The unknown parameters $\lambda_0$, $m_0$, $V_{0,\xi}$ and $V_{0,\epsilon}$ (reminding that the two latter are the variances of $\xi$ and $\epsilon$) belong to $\Theta := [\lambda_-, \lambda_+] \times \Theta_m \times \Theta_{V_\xi} \times \Theta_{V_\epsilon}$ for some known respective intervals $[\lambda_-, \lambda_+]$ and compact intervals $\Theta_m$, $\Theta_{V_\xi}$ and $\Theta_{V_\epsilon}$ included in $[0, \infty)$, with $0 < \lambda_- < \lambda_+ < 1$, and the generating function $f_\nu : x \in [0,1] \mapsto \sum_{s=0}^{\infty} x^s \nu_{s+1}$ belongs to some known class of function $\mathcal{F}$.

(**A5**): There exists a known quantity $K_m > 0$ such that

$$\inf_{(\lambda,\mu,V_\xi,V_\epsilon)\in\Theta,\ f\in\mathcal{F}} |\Xi(\lambda, \mu, V_\xi, V_\epsilon, f)| \geq K_m \tag{5}$$

where $\Xi$ is the function defined on $\Theta \times \mathcal{F}$ by

$$\Xi(\lambda, \mu, V_\xi, V_\epsilon, f) := -\frac{V_\xi m_0 \lambda}{(1-\lambda^2)(1+\lambda)} - \frac{V_\epsilon \lambda}{1-\lambda^2} - \frac{\lambda^2}{1-\lambda^2} f(\lambda) - \frac{1}{1-\lambda^2} f(\lambda^{-1}) \tag{6}$$

where $(\lambda, \mu, V_\xi, V_\epsilon, f) \in \Theta \times \mathcal{F}$.

(**A6**): There exists a known constant $C_Y > 0$ such that the $k$-th moments $\|Y_0\|_k$ of the stationary process, $k = 1, 2$ are less than $C_Y$.

We give a few comments on the two last assumptions. It may be proved that Assumption (**A6**) holds in one of the following situations:

- When $\mathcal{F}$ is the set of power series with non negative coefficents, meaning that the immigration sequence $\{\epsilon,\ n \in \mathbb{Z}\}$ is positively correlated, i.e. $\nu_n \geq 0$ for all $n \geq 1$, in which case $K_m$ is explicit. This is the case when one models the evolution of a disease with constant (increasing or decreasing) trend.
- When $\mathcal{F}$ is included in the set of functions bounded by some constant $M_\mathcal{F}$, in which case we may prove that $K_m$ is explcit and positive if the bound $M_\mathcal{F}$ is small enough.

As to (**A7**), it may be proved that an explicit $C_Y$ may be obtained e.g. if we assume that $\mathbb{E}(\xi^2) \leq 1$.

In order to state the main result of this section, we first introduce some auxiliary functions that will enable us to construct the estimator for the unknown parameters $\lambda_0$ and $m_0$. We first let $\varpi : \mathbb{R} \longrightarrow \mathbb{R}$ be a twice differentiable function such that

$$\varpi(x) \quad \begin{cases} = 1, & x \in [1, +\infty), \\ = 0, & x \in (-\infty, 0], \\ \in [0,1], & x \in [0,1], \end{cases} \tag{7}$$

$$\varpi''(x) = o(x), \quad x \to 0, \tag{8}$$

so that $\varpi'(x) = \mathrm{o}(x^2)$ and $\varpi(x) = \mathrm{o}(x^3)$ as $x \to 0$.

We also need a twice differentiable function $G : \mathbb{R} \longrightarrow \mathbb{R}$ with finite support such that $G(x) = 1$ for $x \in [0, \max(C_Y, C_Y^2)]$. We then let

$$\varpi_k : x \in \mathbb{R} \mapsto \varpi\left(\frac{2}{K_m \lambda_-^k} x\right), \quad k \in \mathbb{N}, \tag{9}$$

$K_m$ being defined in (5), so that $\varpi_k(x)$ is equal to 1 on $\left[2^{-1} K_m \lambda_-^k, \infty\right)$, lies in $[0,1]$ on $\left[0, 2^{-1} K_m \lambda_-^k\right]$ and is 0 on $(-\infty, 0]$. We finally define for all $k \in \mathbb{N} \setminus \{0\}$

$$H_k : x \in \mathbb{R} \;\mapsto\; \frac{1}{k} \varpi_k(|x|) \ln |x|, \tag{10}$$

$$\psi_k : (a,b) \in \mathbb{R}^2 \;\mapsto\; G(a) G(b) H_k(a^2 - b). \tag{11}$$

The following estimator is then defined:

$$\hat{S}_n := \psi_{k_n}(\bar{Y}_n, \bar{Y}_{k_n+1,n}),$$

for a sequence $(k_n)_{n \in \mathbb{N}}$ which is adequately chosen later on. We first state a consistency result.

**Theorem 3** (Quadratic convergence of estimator)**.** Let us suppose that (**A1**), (**A2**), (**A3**), (**A3**)$_\mathbf{2}$, (**A4**), (**A5**) and (**A6**) hold. Let us set $k_n := \lfloor c \ln n \rfloor$ where $c < -\frac{1}{2 \ln \lambda_-}$. The following convergence in quadratic mean holds:

$$\left|\left| \hat{S}_n - \ln \lambda_0 \right|\right|_2 = \mathrm{O}\left(\frac{1}{\ln n}\right) \longrightarrow 0, \quad n \to \infty \tag{12}$$

so that, in particular, $e^{\hat{S}_n}$ converges in probability towards $\lambda_0$ as $n \to \infty$. Furthermore, the estimator defined by $\hat{N}_n := \bar{Y}_n \left(1 - e^{\hat{S}_n}\right)$ converges in probability towards $m_0$ as $n \to \infty$.

The previous result may be refined under slightly stronger assumptions as follows.

**Theorem 4** (Expansion for $\hat{S}_n$)**.** Let us in addition assume here that the covariance of the immigration process satisfies $\nu_h = \mathrm{O}(\zeta^h)$ for some $\zeta < \lambda_-$ (ensuring that (**A2**) holds). Let us set $k_n := \lfloor c \ln n \rfloor$ where $c \in \left(-\frac{1}{2 \ln \zeta}, -\frac{1}{2 \ln \lambda_-}\right)$. Then one has the two terms expansion

$$\hat{S}_n - \ln \lambda_0 = \frac{1}{k_n} \ln \left| \sum_{j=0}^{\infty} \lambda_0^{-j} \chi_j + \lambda_0 (C_1^2 - C_2) \right| + \frac{1}{\sqrt{n} k_n \lambda_0^{k_n}} Z_n \tag{13}$$

where $Z_n$ satisfies $Z_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma)$, $n \to \infty$, for some explicit $\sigma > 0$.

# References


Al-Osh, M. A. and Alzaid, A. A. (1987), First-order integer-valued autoregressive (INAR(1)) process, *Journal of Time Series Analysis*, 8 (3), pp.261–275.

Barczy, M., Basrak, B., Kevei, P., Pap, G. and Planinić, H. (2021), Statistical inference of subcritical strongly stationary Galton-Watson processes with regularly varying immigration, *Stochastic Processes and their Applications*, 132, pp.33–75.

Klimko, L.A. and Nelson, P.I. (1978), On conditional least squares estimation for stochastic processes, *Annals of Statistics*, 6(3), pp.629–642.

Körmendi, K. and Pap, G. (2018), Statistical inference of 2-type critical Galton–Watson processes with immigration, *Statistical Inference for Stochastic Processes*, 21, pp.169–190.

Ispány, M., Körmendi, K. and Pap, G. (2014), Asymptotic behavior of CLS estimators for 2-type doubly symmetric critical Galton–Watson processes with immigration, *Bernoulli*, 20(4), pp.2247–2277.

Venkataraman, K.N. (1982), A time series approach to the study of the simple subcritical Galton–Watson process with immigration, *Advances in Applied Probability*, 14(1), pp.1–20.

Wei, C. Z. and Winnicki, J. (1990), Estimation of the means in the branching process with immigration, *Annals of Statistics*, 18(4), pp.1757–1773.

# Plans d'experiences pour la calibration de code de calcul coûteux

Adama Barry [1,2,3] François Bachoc [1] Clémentine Prieur [2]
Sarah Bouquet [3] Miguel Munoz Zuniga [3]

[1] *Institut de Mathématiques de Toulouse,*
[2] *Université de Grenoble,*
[3] *IFP Énergies Nouvelles*
*adama.barry@math.univ-toulouse.fr*

**Résumé.** Dans l'industrie, le développement de codes de calcul est souvent mis en œuvre pour étudier et analyser des phénomènes ou des systèmes physiques complexes. Certains de ces codes de calculs dépendent de deux types de variables d'entrées : les variables expérimentales ou de contrôle et les paramètres intrinsèques. Ces paramètres sont souvent des constantes physiques et/ou des paramètres de contrôle qui n'ont pas d'interprétation physique. Des valeurs précises doivent être fixées pour ces paramètres par l'ingénieur afin que le code de calcul imite le plus fidèlement possible le phénomène physique d'intérêt. La nature complexe de ces codes de calculs exige une procédure de calibration efficace, dans laquelle les paramètres inconnus sont ajustés pour améliorer l'alignement entre les sorties du code et les observations physiques. La plupart des travaux sur la calibration bayésienne se concentrent sur la construction d'un émulateur du code de calcul en sélectionnant les expériences numériques sans se préoccuper de la qualité des mesures physiques en amont. Étant donné que ces mesures physiques sont limitées par leur coût ou la difficulté de les acquérir, il serait judicieux de les sélectionner pour une calibration plus efficace.

Sur la base du cadre bayésien classique de Kennedy and O'Hagan [2001], nous proposons une stratégie hybride pour la sélection de plans d'expériences physiques et numériques pour la calibration de code de calcul coûteux. Cette stratégie permet une construction précise de l'émulateur de code de calcul, conduisant à une meilleure approximation de la densité a posteriori des paramètres de calibration et, par conséquent, à des résultats de calibration plus précis.

La première étape consiste à sélectionner les expériences physiques. Nous commencerons donc par présenter des critères permettant de mesurer la qualité d'un plan d'expériences physiques. Ces critères peuvent être regroupés en deux catégories : ceux basés sur la matrice d'information Pronzato and Walter [1985], Fedorov [1980] et ceux basés sur la distribution a posteriori exacte. Ces derniers sont mieux adaptées au problème de calibration car elles tiennent compte de la nature non linéaire du code de calcul, de l'incertitude du phénomène physique et des paramètres. Nous présenterons un algorithme d'optimisation pour la résolution du problème d'optimisation de ces critères.
La seconde étape consiste à sélectionner le plan d'expériences numériques. Nous présenterons des critères tirés de la littérature Damblin et al. [2018], Dai and Chien [2018] et ceux que nous proposons. Nos critères s'inspirent du paradigme de réduction séquentielle de l'incertitude

(SUR) Chevalier et al. [2014]. Ils sont basés sur des mesures d'incertitude pour l'objectif de calibration. Pour leur optimisation, qui est coûteuse, nous utiliserons un algorithme glouton qui exploite la procédure de Monte Carlo utilisée dans leur calcul. Une étude comparative sur un cas analytique sera présentée à la fin pour illustrer la performance des différents algorithmes.

**Mots-clés.** Calibration bayésienne, processus gaussien, quantification d'incertitudes, plans d'expériences physiques, plans d'expériences numériques, divergence de Kullback-Leibler.

**Abstract.** In industry, computer codes are often developed to study and analyse complex physical phenomena or systems. Some of these computer codes depend on two types of input variables : experimental or control variables and intrinsic parameters. These parameters are often physical constants and/or control parameters that have no physical interpretation. Precise values must be set for these parameters by the engineer so that the computer code imitates the physical phenomenon of interest as closely as possible. The complex nature of these computer codes requires an efficient calibration procedure, in which the unknown parameters are adjusted to improve the alignment between the code outputs and the physical observations. Most work on Bayesian calibration focuses on building an emulator of the computer code by selecting numerical experiments without regard to the quality of the upstream physical measurements. Given that these physical measurements are limited by their cost or the difficulty of acquiring them, it would make sense to select them for more effective calibration.

The first step is to select the physical experiments. We will therefore begin by presenting the criteria for measuring the quality of a design of physical experiments. These criteria can be grouped into two categories : based on the information matrix (Pronzato and Walter [1985], Fedorov [1980]) and the exact a posteriori distribution. The latter are better suited to the calibration problem because they take into account the non-linear nature of the computer code, the uncertainty of the physical phenomenon, and the uncertainty of the parameters. We will present an optimization algorithm for solving the problem of optimizing these criteria. The second stage consists of selecting the design of numerical experiments. We will present criteria taken from the literature and those that we propose. Our criteria are inspired by the sequential uncertainty reduction (SUR) paradigm Chevalier et al. [2014]. They are based on uncertainty measurements for the calibration objective. For their optimization, which is costly, we will use a greedy algorithm that exploits the Monte Carlo procedure used in their calculation. A comparative study on an analytical case will be presented at the end to illustrate the performance of the different algorithms.

**Keywords.** Bayesian calibration, Gaussian process, uncertainty quantification, designs of physical experiments, designs of numerical experiments, Kullback-Leibler divergence.

# 1 Introduction

Le code de calcul d'intérêt est représenté par une fonction paramétrique dépendant de deux types d'entrées : un vecteur de variables de contrôle désigné par $x \in \mathcal{X} \subset \mathbf{R}^d$ et un vecteur de paramètres $\theta \in \Theta \subset \mathbf{R}^p$ appelés paramètres de calibration.
Sur la base du cadre bayésien classique de Kennedy and O'Hagan [2001] (KOH), la relation entre le code de calcul et le système physique est donnée par le modèle statistique suivant

$$\mathbf{Y}_{obs}(x) = f_{code}(x, \theta_0) + \varepsilon_x, \tag{1}$$

où $\theta_0 \in \Theta$ est la vraie valeur du vecteur des paramètres et $\varepsilon_x \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ est l'erreur de mesure. Nous supposons que la variance $\sigma_\varepsilon^2$ est connue. Le cas échéant, elle peut être considérée comme un paramètre d'incertitude supplémentaire incroporé dans le cadre bayésien ci-dessous.
Notons par $X$ le plan d'expériences physiques et $Y$ les mesures physiques correspondantes. On modélise notre connaissance a priori sur le vecteur de paramètres de calibration par une loi a priori représentée par la densité de probabilité $\pi_0 : \theta \in \Theta \longmapsto \pi_0(\theta) \in \mathbf{R}_+$. La loi a priori est mise a jour, par la règle de Bayes, à l'aide des observations physiques pour donner la densité a posteriori

$$\begin{aligned}
\pi(\theta \mid Y_{obs}) &= \frac{\mathcal{L}(Y_{obs} \mid \theta)\pi_0(\theta)}{Z} \\
&\propto \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\Big[ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} \big(y_i - f_{code}(x^{(i)}, \theta)\big)^2 \Big] \pi_0(\theta) \\
&\propto \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\Big[ -\frac{1}{2\sigma_\varepsilon^2} SS(\theta) \Big] \pi_0(\theta),
\end{aligned} \tag{2}$$

où $\mathcal{L}(Y_{obs} \mid \theta)$ représente la vraisemblance, $Z = \int_\Theta \mathcal{L}(Y_{obs} \mid u)\pi_0(u)du$ la constante de normalisation et $SS(\theta) = \sum_{i=1}^{n} \big(y_i - f_{code}(x^{(i)}, \theta)\big)^2$ la somme des écarts aux carrés.

Pour estimer les paramètres de calibration, nous devons échantillonner la distribution a posteriori (2) par la méthode de Monte-Carlo par chaînes de Markov (MCMC). Cela nécessite un grand nombre d'appels au code de calcul. Ce dernier étant coûteux, cette approche n'est pas réalisable. La solution consiste à approximer la densité a posteriori à l'aide d'un émulateur du code de calcul.

# 2 Émulation par processus gaussien et approximation de la densité a posteriori

On pose un a priori sur le code de calcul comme étant la réalisation d'un processus gaussien

$$\mathrm{Y}_{code} \sim \mathbf{GP}(m_\beta, k_\psi), \tag{3}$$

où $m : u \in \mathcal{X} \times \Theta \longmapsto m(u) \in \mathbf{R}$ est la fonction moyenne a priori et $k_\psi : (u, v) \in (\mathcal{X} \times \Theta)^2 \longmapsto k_\psi(u, v) \in \mathbf{R}$ est la fonction de covariance a priori avec $\beta$ et $\psi$ le vecteur des paramètres de régressions et le vecteur des hyperparamètres.

Supposons que l'on dispose de $D_M = \Big((x_1,\theta_1),\ldots,(x_M,\theta_M)\Big)^T$ le plan d'expériences numériques et $f_{code}(D_M) = \Big(f_{code}(x_1,\theta_1),\ldots,f_{code}(x_M,\theta_M)\Big)^T$ les observations numériques correspondantes. Le processus gaussien conditionné aux observations numériques reste également gaussien :

$$Y^M_{code} := \Big[Y_{code} \mid Y_{code}(D_M) = f_{code}(D_M)\Big] \sim \mathbf{GP}\Big(\mu_M, k_M\Big), \tag{4}$$

où $\mu_M$ et $k_M$ representent la fonction moyenne a posteriori et la fonction de covariance a posteriori dont les expressions suivent :

$$\mu_M(\upsilon) = m_\beta(\upsilon) - k(\upsilon, D_M)\big[k(D_M)\big]^{-1}\big[f_{code}(D_M) - m_\beta(D_M)\big], \tag{5}$$

$$k_M(\upsilon,\upsilon') = k(\upsilon,\upsilon') - k(\upsilon, D_M)\big[k(D_M)\big]^{-1}k(D_M,\upsilon'), \tag{6}$$

et

$$m_\beta(D_M) = \Big(m_\beta(x_i,\theta_i)\Big)_{i=1,\ldots,M} \quad , \quad k(\upsilon, D_M) = \Big(k\big(\upsilon,(x_i,\theta_i)\big)\Big)_{i=1,\ldots,M},$$
$$k(D_M,\upsilon') = \Big(k\big((x_i,\theta_i),\upsilon'\big)\Big)_{i=1,\ldots,M} \quad , \quad k(D_M) = \Big(k\big((x_i,\theta_i),(x_j,\theta_j)\big)\Big)_{i,j=1,\ldots,M}.$$

Les paramètres du modèle de processus gaussien $(\beta,\psi)$ sont omis dans les notations par souci de simplicité. Ils peuvent être estimés à l'aide des observations numériques par la technique de modularisation (voir Damblin et al. [2018]). Une hypothèse supplémentaire d'indépendance entre la loi a priori du vecteur des paramètres et celle du code de calcul conduit à l'approximation de la densité a posteriori donnée comme suit :

$$\begin{aligned}\pi(\theta \mid Y_{obs}, f_{code}(D_M)) &= \frac{\mathcal{L}^c(Y_{obs} \mid f_{code}(D_M),\theta)\pi(\theta)}{\int_\Theta \mathcal{L}^c(Y_{obs} \mid f_{code}(D_M),u)\pi(\theta')du} \\ &\propto \mathcal{L}^c(Y_{obs} \mid f_{code}(D_M),\theta)\pi(\theta),\end{aligned} \tag{7}$$

où

$$\mathcal{L}^c\big(Y_{obs} \mid f_{code}(D_M,\theta)\big) = \frac{1}{(2\pi)^{n/2} \mid W_M(\theta) \mid^{1/2}} \times \exp\Big[-\frac{1}{2}\big(Y_{obs} - \mu_M(X_{obs},\theta)\big)^T\big[W_M(\theta)\big]^{-1}\big(Y_{obs} - \mu_M(X_{obs},\theta)\big)\Big]$$

est la vraisemblance conditionnée aux observations numériques, avec $W_M(\theta) = \sigma^2_\varepsilon I_n + k_M\big((X_{obs},\theta),(X_{obs},\theta)\big)$.

De l'équation (7), nous pouvons noter que la qualité de l'approximation dépend de la qualité de l'émulateur qui à son tour dépend des observations numériques donc du plan d'expériences numériques. De plus la qualité de la densité a posteriori que l'on cherche à approximer (2) dépend de la qualité des observations physiques. D'où la nécessité de sélectionner avec soin le plan d'expériences physiques et le plan d'expériences numériques pour une calibration efficace.

# 3 Algorithme hybride pour la calibration de code de calcul

L'algorithme hybride que nous proposons pour la calibration se déclinent en cinq étapes principales :

1. Construire un émulateur de processus gaussien initial à l'aide d'un plan d'expériences numériques $D_{M_0}$ et les observations numériques correspondantes $f(D_{M_0})$.
2. Utiliser l'émulateur pour sélectionner le plan d'expériences physiques

$$X_{obs} \in \arg\max \mathbf{C}(X), \tag{8}$$

   où $C$ est un critère de sélection qui mesure la quantité d'information contenue dans un plan d'expériences physiques.
3. Effectuer les mesures sur le terrain et recueillir les observations physiques $Y_{obs}$.
4. Pour $k = M_0, \ldots, M$ :
   — Sélectionner en deux étapes :

$$\theta_{k+1} \in \arg\max_{\Theta} \alpha_k(\theta) \tag{9}$$

$$x_{k+1} \in \arg\max_{X_{obs}} \beta_k(x, \theta_{k+1}), \tag{10}$$

   où $\alpha_k$ et $\beta_k$ sont des fonctions d'acquisitions.
   — Enrichir le plan d'expériences numériques $D_{k+1} = D_k \cup \{(x_{k+1}, \theta_{k+1})\}$ et les observations numériques $f_{code}(D_{k+1}) = \big(f_{code}(D_{k+1})^T, f_{code}(x_{k+1}, \theta_{k+1})\big)^T$.
   — Mettre à jour l'émulateur de processus gaussien.
5. Approximation de la distribution a posteriori, échantillonnage MCMC et estimation des intervalles de crédibilité des paramètres de calibration.

L'exposé se concentrera sur les critères de la littérature basés sur la matrice d'information (Pronzato and Walter [1985], Fedorov [1980]) et ceux que nous proposons qui utilisent la distribution a posteriori pour mesurer la qualité d'un plan d'expériences physiques.
Pour la sélection séquentielle de plans d'expériences numériques, nous présenterons les stratégies de Damblin et al. [2018], Dai and Chien [2018] et Gardner et al. [2019] et celles que nous proposons inspirées du paradigme de réduction séquentielle d'incertitudes (Chevalier et al. [2014]). Des algorithmes d'optimisation adaptés aux problèmes (8), (9) et (10) seront également présentés. Et enfin des illustrations et des résultats des performances des algorithmes sur des cas test analytiques seront présentés.

# Références


Clément Chevalier, David Ginsbourger, Victor Picheny, Julien Bect, Emmanuel Vazquez, and Yann Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4) :455–465, 2014. ISSN 00401706, 15372723. URL http://www.jstor.org/stable/24587032.

Xiaowu Dai and Peter Chien. Another look at statistical calibration : A non-asymptotic theory and prediction-oriented optimality, 2018. URL https://arxiv.org/abs/1802.00021.

Guillaume Damblin, Pierre Barbillon, Merlin Keller, Alberto Pasanisi, and Éric Parent. Adaptive numerical designs for the calibration of computer codes. *SIAM/ASA Journal on Uncertainty Quantification*, 6(1) :151–179, Jan 2018. ISSN 2166-2525. doi : 10.1137/15m1033162. URL http://dx.doi.org/10.1137/15M1033162.

Valery V. Fedorov. Convex design theory 1. *Statistics*, 11 :21–43, 1980.

Paul Gardner, Charles Lord, and Robert Barthorpe. Sequential bayesian history matching for model calibration. 05 2019. doi : 10.1115/VVS2019-5149.

Marc Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, 63 :425–464, 02 2001. doi : 10.1111/1467-9868.00294.

Luc Pronzato and Eric Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1) :103–120, 1985. ISSN 0025-5564. doi : https ://doi.org/10.1016/0025-5564(85)90068-9. URL https://www.sciencedirect.com/science/article/pii/0025556485900689.

# Spatio-temporal weather generator for the temperature over France

Caroline Cognot [1&2] & Liliane Bel [2] & Sylvie Parey [1] & David Métivier [3]

[1] *EDF R&D, 91120, Palaiseau, France - caroline.cognot@edf.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France*
[3] *UMR MISTEA, 34000, Montpellier, France*

**Résumé.** Les générateurs de temps sont des simulateurs qui permettent de reproduire la variabilité climatique et générer un grand nombre de situations pour des variables météorologiques selon un modèle statistique ajusté sur les observations. La plupart des générateurs de temps fournissent des simulations pour une ou plusieurs variables météorologiques site par site. Dans ce travail, nous nous proposons de concevoir un générateur spatialisé pour une seule variable météorologique. Nous nous focalisons sur la température. La moyenne et la variance sont décomposées chacune en une tendance et une saisonnalité. Les tendances sont modélisées par des fonctions non paramétriques, les saisonnalités par des polynômes trigonométriques. Chaque fonction est ajustée site par site, puis étendue à tout l'espace par krigeage pour les saisonnalités et par distance inverse pour la partie tendance. La partie stochastique est modélisée par un champ gaussien avec une fonction de corrélation spatio-temporelle non séparable. Ces deux étapes permettent de simuler sur une grille, sur n'importe quelle période de temps. La validation du modèle a été réalisée à la fois sur le jeu de données des stations et sur la grille. En calculant plusieurs indicateurs liés à la structure spatiale, à la structure temporelle ou aux extrêmes, on constate que le générateur fournit des simulations adéquates et permet la génération de séries de températures spatialement cohérentes, à l'exception toutefois des extrêmes élevés, ce qui était attendu, les processus gaussiens étant connus pour être peu performants sur cet aspect.

**Mots-clés.** Générateur de temps, processus spatio-temporel, processus Gaussiens

**Abstract.**

Weather generators are simulators useful to reproduce climate variability and generate a great number of situations for meteorological variables according to a statistical model fitted on observations. Most weather generators provide simulations for one or several meteorological variables site by site, we aim in this work to design a spatial weather generator for one meteorological variable. We focus on temperature. The mean and variance are each modeled as a trend and a seasonality function. Trends are non-parametric functions while seasonality parts are written as trigonometric polynomials. They are fitted separately on each site and then extended to the whole domain by kriging for the seasonality coefficient and by inverse distance weighting for the trends. The stochastic part is modeled as a Gaussian field with a non-separable spatio-temporal correlation function. These two steps allow for simulation on a grid, over any time period. The validation of the model was conducted both on the stations dataset and the grid. By computing several indicators related to the spatial structure, the

temporal structure or extremes, it is found that the generator provides adequate simulations and allows the generation of spatially coherent temperature series, except for high extremes, which was to be expected as Gaussian processes are known to perform poorly in this aspect.

**Keywords.** Weather generator, Spatio-temporal process, Gaussian Processes

# 1 Introduction

Weather generators consider the observations of a meteorological variable to be one of the infinite possible trajectories of a stochastic model. The main objective is to allow the simulation of many plausible sequences of a meteorological variable, that share common statistical properties with the observations. The parameters are estimated on the observations. Simulations are fast and allow for a more complete sampling of the climate variables than physical models. Since the first weather generator of Richardson [1981] many others have been proposed using various statistical tools, for univariate and multivariate series. In the case of multisite temperature-focused models, Dubrovsky et al. [2020] propose an auto-regressive model with a high number of matrix parameters to estimate, and Sparks et al. [2018] use Empirical Orthogonal Functions for dimension reduction. These approaches however do not take advantage of the spatial structure of the data. A good option is to turn to geostatistics. The main improvement we can get from these tools is the use of spatial information in the dependencies, with correlation functions that involve spatial distances. Many models use this concept using one or several Gaussian fields with a chosen covariance function. They are combined if necessary with an anamorphosis function, as in Kleiber et al. [2012], Verdin et al. [2015], or Sparks et al. [2018], in order to be used on non-Gaussian data, mainly precipitations. These Gaussian fields can have varying complexity, with spatio-temporal covariance function ranging from the simple $C(h,t) = \exp(-||h||/A(t))$ with $h$ the distance between spatial points and $A$ a scale function in Kleiber et al. [2012] to non-separable covariance functions in Bourotte et al. [2016],Allard et al. [2022].

In this work, we focused on building a spatio-temporal Stochastic Weather Generator (SWG) for the temperature. The model is built in two steps: first, the mean and variance are represented as a sum of trend and seasonality functions. Trends are depicted as non-parametric functions, whereas the seasonality components are expressed using trigonometric polynomials. Then, a Gaussian field model is fitted on the residuals.

# 2 Methods

## 2.1 Preprocessing of the temperature

**Single-site decomposition :** Let $X(s,t)$ be the temperature at site $s$ and time $t$, with $s \in \{s_1, ..., s_{n_S}\} \subset D$, and $t \in \{1, .., n_T\}$. At each station, we use the decomposition of Hoang [2010] to separate the mean and variance into trend and seasonality components.

The decomposition writes

$$X(s,t) = T_m(s,t) + S_m(s,t) + T_\sigma(s,t)S_\sigma(s,t)Z(s,t), \tag{1}$$

where $T_m$ is the trend in mean, $S_m$ is the seasonality in mean, $T_\sigma$ is the trend in standard-deviation and $S_\sigma$ is the seasonality in standard-deviation. $Z$ is the residual signal, carrying the random information.

The seasonality terms are modeled by trigonometric polynomials of the following form:

$$S(s,t) = \beta_1(s) + \sum_{i=1}^{d} \left[ \beta_{2i}(s) \cos\left(\frac{2i\pi t}{365}\right) + \beta_{2i+1}(s) \sin\left(\frac{2i\pi t}{365}\right) \right]. \tag{2}$$

For each site $s$, the different components in Eq. (1) are obtained iteratively using LOESS (LOcally Estimated Scatterplot Smoothing) regression for the trend terms $T_m$, $T_\sigma$ and linear regression for the trigonometric polynomial coefficients $\boldsymbol{\beta}_m(s), \boldsymbol{\beta}_\sigma(s)$ involved in $S_m, S_\sigma$. The LOESS regression gives a non-parametric form to the trends, avoiding the over-simplification of linear trends. This model takes into account the seasonal cycle of the variance, which is often neglected in most weather generators.

**Spatial extension:** In order to use the model anywhere in space, the trends $T_m, T_\sigma$ and the coefficients of the seasonality $\boldsymbol{\beta}_m(s), \boldsymbol{\beta}_\sigma(s)$ have to be extended. The trends are assumed to slowly vary in time and are interpolated using inverse distance weighting. For the seasonality coefficients, we use ordinary kriging.

## 2.2 A geostatistics model

Once the coefficients of the decomposition are estimated, we are left with centered residuals $Z$ of variance 1 when averaged over a year. They are spatially and temporally correlated, and in a sense, represent how the weather at time $t$, site $s$ deviates from a generic value. Our approach for the residuals $Z(s,t)$ makes use of geostatistics to include geographical proximity in the dependence structure, and allow simulation at any new point in space.

We suppose $Z(.)$ to be a Gaussian stationary process, with mean 0. Under these two assumptions, it is entirely described by its covariance function $C(h,u)$ where $h$ is a space lag and $u$ is a time lag. We choose to use the Gneiting-Matern covariance model (16) of Gneiting [2002], a non-separable model that allows interaction between space and time,

$$C(h,u) = \frac{\sigma^2}{(\alpha u^{2a}+1)^\tau} \mathcal{M}\left( \frac{h}{\sqrt{\alpha u^{2a}+1}^{b}}; r; \nu \right), \tag{3}$$

where $\mathcal{M}(d,r,\nu)$ is the value of the Matern covariance function at distance $d$, with smoothness parameter $\nu$ and scale parameter $r$.

**Simulation** When working with spatiotemporal grids of limited size, simulating a Gaussian process with the covariance model in Eq.(3) is straightforward, using standard simulation procedures such as Cholesky decomposition for instance. When working with finer grids and long periods of time, this is no longer computationally feasible. We instead choose to simulate sequentially in time making the values at time $t$ only depending on the $l$ previous time steps, with $l$ coherent with the temporal range parameter that is estimated. This is made possible by the Gaussian properties of the model allowing easy simulation conditionally to the previous time steps.

## 2.3 Validation

To validate such a model, we choose to look at various indicators, such as the annual cycle in mean and variance at each point or the pairwise correlations between pairs of sites. The main idea is to compute many simulations, obtain an empirical distribution of these indicators, and compare them to the observed values. The model is adequate if the observed values are in the range of the simulations.

**Pairwise correlations between sites** For each pair of sites $i, j$, we compute for the observations and 100 simulations the pairwise empirical correlations of the temperature $X$.

**Pairwise conditional threshold exceedance** We are interested in how extremes in one location are related to the same extreme in another location, and we look at pairwise conditional threshold exceedances for high (resp. low) quantiles.

For each station $i$, define the probability $\alpha$ and quantile $q_\alpha(i)$ such that $\alpha = P(X_i > q_\alpha(i))$ (resp. $x = P(X_i < q_\alpha(i))$) in the observations. Then, for every pair of stations $i, j$, define

$$p^\alpha_{i,j} = P(X_i > q_\alpha(i) | X_j > q_\alpha(j)) \tag{4}$$

$$\hat{p}^\alpha_{i,j} = \frac{\sum_{t=1}^{N_t} \mathbb{1}_{X_i > q_\alpha(i) \cap X_j > q_\alpha(j)}}{\sum_{t=1}^{N_t} \mathbb{1}_{X_j > q_\alpha(j)}}. \tag{5}$$

With inverted signs in the case of low quantiles. The values of $\hat{p}^x_{i,j}$ for all $i, j$ in the observations are then compared with the values from the simulations.

**Temporal correlation** As we wish to be able to represent well the temporal structure of the data, we also look at the temporal correlation function. In particular, we compare the lagged correlation in the full covariance simulation and in the simulations where we use only 10 previous days to simulate time $t$.

# 3 Results

## 3.1 Parameter estimation

The data used is a set of 41 ECA&D (Klein Tank [2002]) stations in France. The trends and seasonality are estimated on 61 years ( 1960-2020) and the covariance parameters are estimated on 31 years (1985-2015). The covariance parameters for (3) are computed with a composite pairwise likelihood estimation using R package GeoModels (Bevilacqua et al. [2018]) for each extended season each year (extended winter: September- October - November - December - January- February; extended summer: March - April -May-June-July-August). This separation is both for the sake of computational efficiency and to account for differences between the seasons observed during an initial study. This leads to 30 sets for winter and 31 for summer (the first and last winter are not complete when working on data that start from 1st January), see table 1 for the median and standard deviation of the found parameters. The final set of parameters is chosen as the median over all years for each of the parameters.

| Parameter | Winter | | Summer | |
|---|---|---|---|---|
| | median | sd | median | sd |
| $\sigma^2$ | 1.0199e+00 | 1.252e-01 | 9.713e-01 | 1.474e-01 |
| $\tau$ | 1.006e+00 | 8.715e-01 | 1.142e+00 | 2.606e-01 |
| $\nu$ | 8.662e-01 | 1.750e-01 | 6.834e-01 | 1.282e-01 |
| $2a$ | 1.507e+00 | 2.405e-01 | 1.619e+00 | 2.085e-01 |
| $\frac{r}{1000}$ | 1.000e+00 | 1.173e-03 | 9.997e-01 | 1.026e-03 |
| $\frac{1}{\alpha}$ | 3.118e+00 | 1.082e+00 | 2.871e+00 | 4.072e-01 |
| nugget | 9.631e-02 | 2.297e-02 | 1.009e-01 | 2.285e-02 |

Table 1: Estimated covariance parameters

We find some of the parameters to be quite different between the two seasons. Notably, the smoothing Matérn parameter and the variance are higher in winter than in summer. The temporal scale parameter tends to be higher in the winter but the results are also a lot more dispersed than for the summer.

We also find that seasonal coefficients vary with the geographic coordinates. Compared to a simple longitude/latitude regression, kriging the values of the seasonal coefficients allows to highlight regional characteristics of the French climate. As seen in Figure 1 showing the kriging for the mean coefficients, we can retrieve the regions with oceanic, continental, and Mediterranean climates.

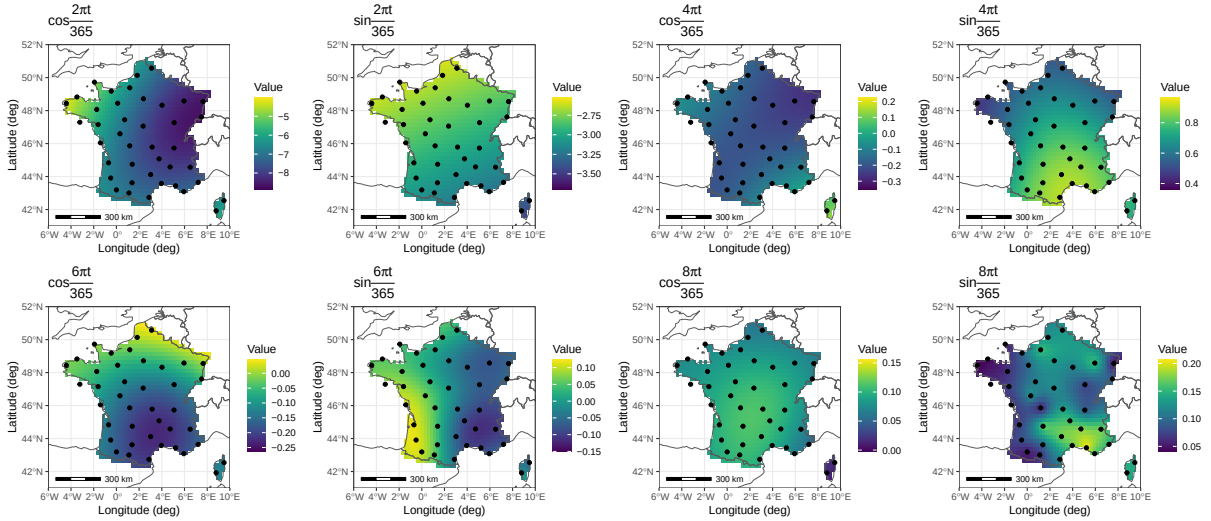



Figure 1: Kriging the coefficients of $S_m$.

## 3.2 Simulation at the fitting points

Temperature series are simulated 100 times at each of the 41 fitting data points in three ways. First, a 31 years simulation is obtained by first simulating 31 summers and 31 winters from the fitted covariance model in Eq. (3) and then adding the deterministic parts in Eq. (1). Then, we simulate from Eq. (3) using 10 days conditional. Finally, we simulate using only a temporal covariance function, to evaluate the gain of the spatial part.

We find that although the spatial correlation is necessarily embedded in the removed deterministic parts, the spatial structure of the residuals is still important to correctly reproduce the spatial correlation of temperature. In figure 2 (left), we plot all the pairwise correlations between stations and compare them to the values computed from the observations. We expect a good model to give points close to the 1:1 line in black. Compared to simulating from a non-spatial model (blue dots), the pairwise correlations between stations are indeed improved by our model (red dots for simulations with full covariance, almost entirely covered by the green dots from the 10 days conditional simulation). Figure 2 (right) shows, for every pair of station $i, j$, the probability of station $i$ exceeding a quantile, given station $j$ does. Again, we expect points close to the 1:1 line, which is not entirely true, especially for very low (q1) or very high (q99) quantiles. This probably comes from the fact that Gaussian fields are not able to produce extremal dependence. However, there is definitely an improvement in using a spatiotemporal model instead of a purely temporal one.

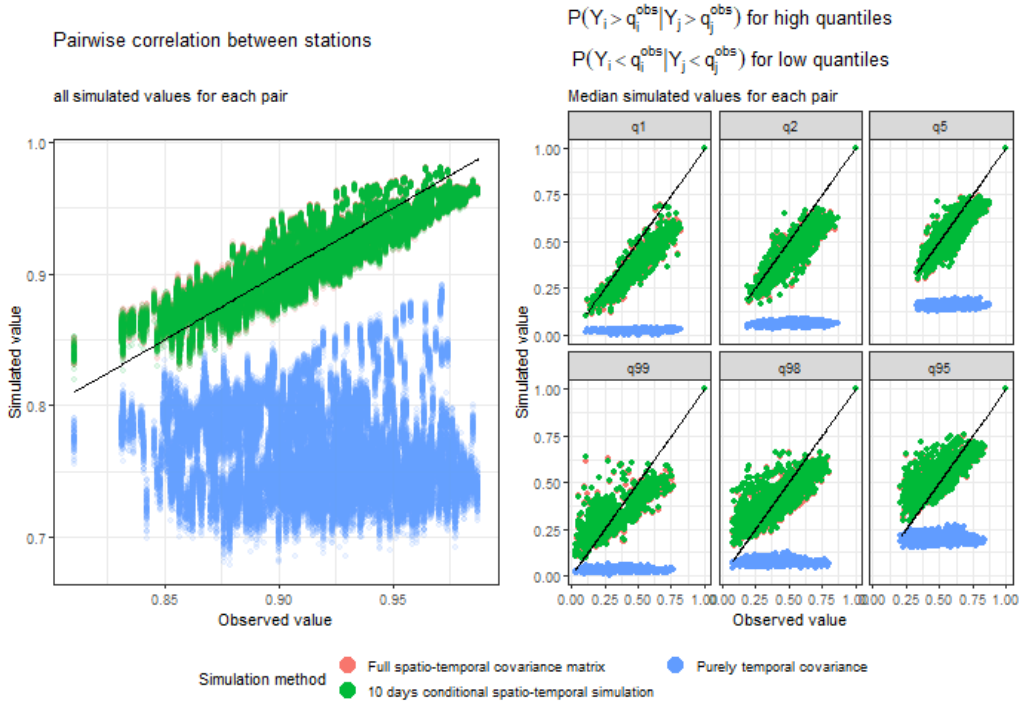



Figure 2: Pairwise correlation between stations (left) and extremal dependence (right).

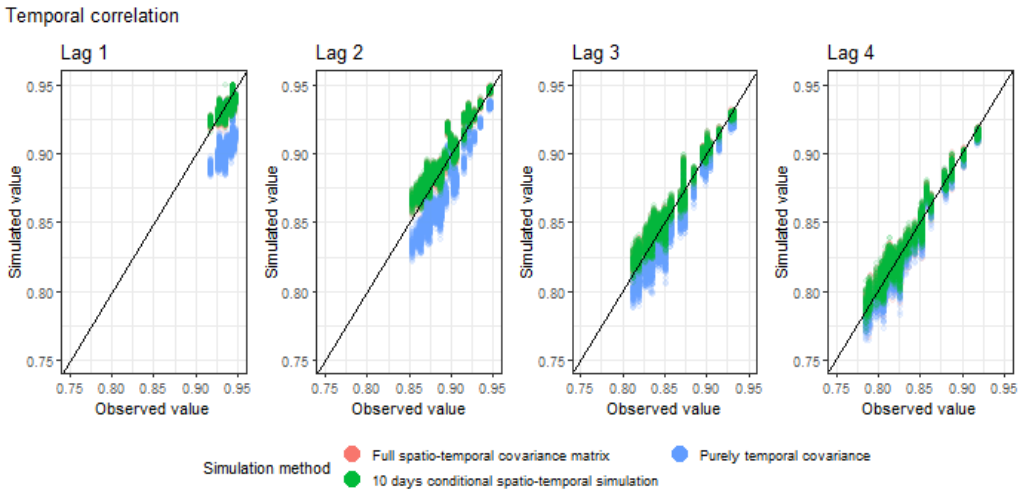



Figure 3: Temporal correlations at each of the 41 stations

The lagged temporal correlations, in figure 3, are also quite well represented in the simulations.

### 3.3 Simulation on a regular grid

We make simulations on a grid using the kriged coefficients in 1, the interpolated trends and residuals simulated from Eq. (3) using the 10 days conditional. Two simulations for the same period in June 2005 are represented in figure 4 .

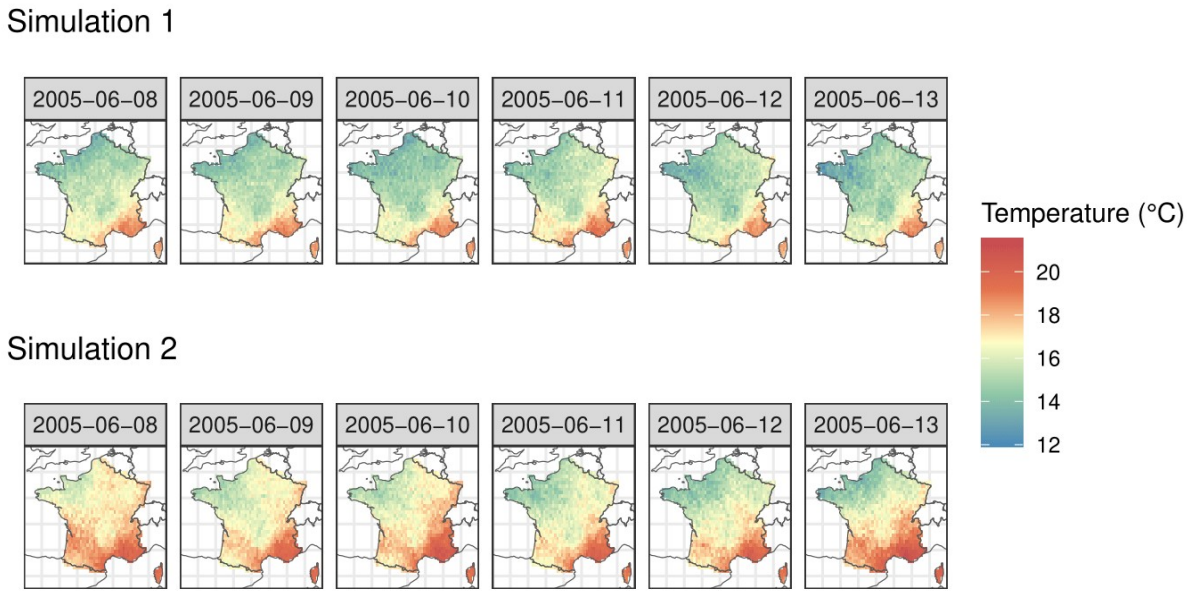



Figure 4: Two simulations, equivalent to the same period in June 2005

The simulated temperature values depend both on the deterministic parts (trends and seasonality) and the stochastic part. We see on both of the 6-day sequences that the two simulations give different results: the simulated temperature is a lot warmer in the southwest in the second simulation than in the first.

## 4 Conclusion and perspectives

In this work, we construct a spatio-temporal weather generator for the temperature. It can be used to simulate on grids over any period of time, provided trends are available. Trends in mean and standard deviation can be derived from climatic studies while supposing the stochastic structure to be stationary.

Our simulations will be compared with the E-OBS dataset. The E-OBS dataset is a gridded set of observations, obtained from interpolating from the ECA&D dataset, the same we used in this study. Because of that, it can be interesting to see whether our model is consistent with it.

We will investigate other simulation methods for Gaussian processes in order to be able to handle large datasets and several meteorological variables.

# References


D. Allard, L. Clarotto, and X. Emery. Fully nonseparable gneiting covariance functions for multivariate space–time data. Spatial Statistics, 52:100706, 2022.

M. Bevilacqua, V. Morales-Oñate, and C. Caamaño-Carrillo. GeoModels: Procedures for Gaussian and Non Gaussian Geostatistical (Large) Data Analysis, 2018. URL https://vmoprojs.github.io/GeoModels-page/. R package version 1.0.0.

M. Bourotte, D. Allard, and E. Porcu. A flexible class of non-separable cross-covariance functions for multivariate space–time data. Spatial Statistics, 18:125–146, 2016.

M. Dubrovsky, R. Huth, H. Dabhi, and M. W. Rotach. Parametric gridded weather generator for use in present and future climates: focus on spatial temperature characteristics. Theoretical and Applied Climatology, 139:1031–1044, 2020.

T. Gneiting. Nonseparable, stationary covariance functions for space–time data. Journal of the American Statistical Association, 97(458):590–600, 2002.

T. T. H. Hoang. Modeling of non-stationnary, non-linear time series - Application to the definition of trends on mean, variability and extremes of air temperatures in Europe. PhD thesis, Université Paris Sud-Paris XI, 2010.

W. Kleiber, R. W. Katz, and B. Rajagopalan. Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes. Water Resources Research, 48(1), 2012.

A. a. C. Klein Tank. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453, 2002.

C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. Water resources research, 17(1):182–190, 1981.

N. J. Sparks, S. R. Hardwick, M. Schmid, and R. Toumi. Image: a multivariate multi-site stochastic weather generator for european weather and climate. Stochastic environmental research and risk assessment, 32:771–784, 2018.

A. Verdin, B. Rajagopalan, W. Kleiber, and R. W. Katz. Coupled stochastic weather generation using spatial and generalized linear models. Stochastic Environmental Research and Risk Assessment, 29:347–356, 2015.

# Session groupe Jeunes Statisticiens

# Sexisme ordinaire, violences sexistes et sexuelles, biais de genre
# Quel est le constat aujourd'hui dans la recherche académique en France ?

Perrine Lacroix [1], Arthur Leroy [2], Iqraa Meah [3], Antonio Ocello [4], Eloi Tanguy [5], Charlotte Voinot [6] & Margaux Zaffran [7]

[1] *Laboratoire de Biologie et Modélisation de la Cellule, ENS de Lyon, Lyon, France, perrine.lacroix@ens-lyon.fr*
[2] *Department of Computer Science, The University of Manchester, Manchester, United Kingdom, arthur.leroy.pro@gmail.com*
[3] *METHODS, CRESS, Paris, France, iqraa.meah@inserm.fr*
[4] *CMAP, École Polytechnique, Palaiseau, France, antonio.ocello@polytechnique.edu*
[5] *MAP5, Université Paris-Cité, Paris, France, eloi.tanguy@u-paris.fr*
[6] *SANOFI, France – PreMeDICaL INRIA, Montpellier, France, Charlotte.Voinot@sanofi.com*
[7] *PreMeDICaL INRIA, Montpellier, France – CMAP, École Polytechnique, Palaiseau, France, margaux.zaffran@inria.fr*

**Résumé.**

Le Groupe Jeunes de la SFdS s'engage depuis plusieurs années maintenant à créer un espace de discussion nécessaire pour sensibiliser la communauté académique à des sujets difficiles, comme la santé mentale des jeunes chercheur.e.s ou l'éthique dans la recherche. La session jeune prend place chaque année aux Journées de Statistique et a pour but de promouvoir l'échange et de proposer un cadre de réflexion sain et constructif autour d'une thématique choisie. Cette année, nous avons choisi de construire une session sur le thème du sexisme, des violences et biais de genre dans le monde académique. Le sexisme, les biais de genre ou plus gravement les Violences Sexistes et Sexuelles (VSS) sont malheureusement encore trop présents dans le milieu académique. Un des premiers objectifs de cet évènement sera notamment d'actualiser le constat sur les inégalités de genres et les VSS en 2024 dans le milieu académique en France. L'évolution même des carrières académiques souffre encore de fort biais genrés, à l'heure où la France compte seulement 28% de femmes professeur.e.s des universités (alors qu'elles représentent 45% des maître.sse.s de conférence, et 49% des docteur.e.s). La session sera construire de manière à mettre en place un dialogue entre le public et des intervenant.e.s ayant travaillé ou milité sur ces questions. Un autre objectif supplémentaire sera de partager des informations à caractère préventif, telles que les notions de bases pour intervenir en cas de VSS. En participant à cette session, vous contribuerez non seulement à l'avancement du débat sur ces questions cruciales, mais vous vous doterez également d'outils et de connaissances pour promouvoir un environnement académique plus inclusif et respectueux pour tous et toutes.

**Mots-clés.** Sexisme, Violences sexistes et sexuelles, Biais de genre

**Abstract.** The SFdS Youth Group has been committed for several years now to creating a necessary space for discussion to raise awareness within the academic community about challenging topics, such as the mental health of young researchers or ethics in research. The youth session takes place annually at the Journées de Statistique to promote exchange and offer a framework for healthy and constructive reflection on a chosen theme. This year, we have chosen to build a dedicated session on sexism, violence, and gender bias in the academic world. Sexism, gender bias, or, more seriously, Sexist and Sexual Violence (SSV), unfortunately, are still too prevalent in the academic environment. One of the main objectives of this session would be to update the assessment of gender inequalities and SSV in the academic environment in France in 2024. Even the evolution of academic careers still suffers from persistent gender biases, as only 28% of University professors are women in France nowadays (while they represent 45% of the lecturers and 49% of all doctors). The session will be structured to facilitate a dialogue among the audience and possibly specialist speakers on these issues. Another objective of the session will be to provide basic knowledge for intervening in cases of SSV. By participating in this session, you will not only contribute to advancing the debate on these crucial issues but also equip yourself with tools and knowledge to promote a more inclusive and respectful academic environment for everyone.

**Keywords.** Sexism, Sexist and sexual violence, Gender bias

# Session groupe Statistique et Sport

# A data-driven approach to select the best compositions of a wheelchair basketball team

Gabriel Calvo [1] & Carmen Armero [2] & Bernd Grimm [3] & Christophe Ley [4]

[1] *Department of Statistics and Operations Research, Universitat de València, Spain*
*gabriel.calvo@uv.es*
[2] *Department of Statistics and Operations Research, Universitat de València, Spain*
*carmen.armero@uv.es*
[3] *Department of Precision Health, Luxembourg Institute of Health, Luxembourg*
*bp.grimm@gmail.com*
[4] *Department of Mathematics, University of Luxembourg, Luxembourg*
*christophe.ley@uni.lu*

**Abstract.** This paper explores optimising team line-ups in wheelchair basketball, a sport governed by the International Wheelchair Basketball Federation (IWBF) and its Player Classification Points System (PCPS). It focuses on evaluating players' performances using metrics like the efficiency rating (EFF), performance index rating (PIR), and Win score. The study employs Bayesian longitudinal modelling alongside an optimisation programming respecting game constraints, including players' gender, classification, and match location, to predict the best team compositions. This methodology is applied to the Doneck Dolphins Trier team from the German Rollstuhlbasketball-Bundesliga (RBB), highlighting the effectiveness of these models in preparing for the 2022-2023 season matches. The study demonstrates a novel approach to team optimisation in wheelchair basketball, considering both player performance and game regulations.

**Keywords.** Bayesian statistics, Longitudinal data, Sport analytics

## 1 Introduction

The present work is based on Calvo et al. (2023) and it is motivated by sports, specifically wheelchair basketball, regulated by the International Wheelchair Basketball Federation (IWBF). This sport employs a Player Classification Points System (PCPS) that assigns a rating to each player ranging from 1 to 4.5, reflecting the spectrum from the player with the least physical capacity to the one in excellent condition. The sum of the scores of the five players on the court must never exceed 14. We introduce an innovative methodology for selecting the best line-ups for a wheelchair basketball team for a future game, based on performance metrics of different players in previous games. We combine Bayesian modeling, analysing players' performance data throughout the season, with an integer linear programming scenario to determine line-ups that maximise the team's performance subject to game constraints. One key feature of this proposal is the incorporation of uncertainty, introducing variability into the optimal solution.

# 2 Methodology and application

## 2.1 General framework

The methodology in this work is general and consists of three stages: 1) Approximating the posterior distribution of the parameters of the Bayesian longitudinal model that analyses players' performance metrics. 2) Computing predictions for a new game using the posterior predictive distribution associated to each player. 3) Obtaining the posterior distribution of the solutions to the integer linear programming optimisation problem that seeks to maximise the sum of predictions under certain constraints.

## 2.2 Methodology

In this study, we model the performance of players in basketball games using a longitudinal variable $y_{ij}$, which represents the performance of player $i$ in game $j$. The sampling model is a mixed linear model:

$$\begin{aligned}&(y_{ij} \mid \boldsymbol{\theta}, \boldsymbol{b}) \sim \mathrm{N}(\mu_{ij}, \sigma^2),\\ &(\mu_{ij} \mid \boldsymbol{\theta}, \boldsymbol{b}) = \beta_0 + b_{0i} + b_{0j} + \beta_W I_W(i) + \beta_C C_i + \beta_H I_H(j) + (\beta_1 + b_{1i})j,\end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{b}$ include model parameters and random effects, respectively. Parameters consist of the measurement error standard deviation $\sigma$, common coefficients ($\beta_0$, $\beta_W$, $\beta_C$, $\beta_H$, and $\beta_1$), and the standard deviations of random effects ($\sigma_{0b}$, $\sigma_{0m}$, $\sigma_{1b}$) which are assumed to be normally distributed. Covariates include a gender indicator $I_W(i)$, the functional classification $C_i$ of each player $i$, and a home indicator $I_H(j)$ for each match $j$.

For Bayesian modelling, non-informative and independent prior distributions are chosen. Wide normal distributions are used for coefficients, while uniform distributions are selected for standard deviation parameters.

The study further explores the prediction of player performance in future games using the posterior predictive distribution:

$$f(y_i^{(pre)} \mid \mathcal{D}) = \int f(y_i^{(pre)} \mid \boldsymbol{\theta}, \boldsymbol{b})\, \pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}). \tag{2}$$

Finally, the study proposes a stochastic integer linear programming optimisation problem to select the optimal team composition, maximising combined team performance while adhering to constraints. The objective function to be maximised is:

$$\max \sum_{i=1}^{N} z_i y_i^{(pre)}, \tag{3}$$

subject to player inclusion in the team lineup and functional classification constraints, ensuring compliance with the IWBF's rules.

## 2.3 Application

To illustrate this procedure, data from the 18 games in the 2022-2023 season of the Doneck Dolphins Trier wheelchair basketball team in the German 1st division are analysed. We followed nine players throughout the 18 games of the season, three of whom are female, reflecting the mixed-gender nature of the competition. All player information is summarised in Table 1. Moreover, three individual performance metrics are considered: player efficiency (EFF), performance index rating (PIR), and Win Score. The Bayesian longitudinal model is separately fitted for each metric. Player Dirk Passiwan stands out in the team, with somewhat irregular but exceptionally good performance. Due to the constraints of the Player Classification Points System (PCPS), high individual performance does not necessarily always lead to inclusion in the line-up.

Table 1: Name, functional classification value and sex of the players of the Doneck Dolphins Trier team who play more than 40 minutes during the 2022-2023 season.

| Player | Classification | Sex | Index |
|---|---|---|---|
| Annabel Breuer | 1.5 | Woman | 1 |
| Correy Rossi | 2 | Man | 2 |
| Dejon Green | 3.5 | Man | 3 |
| Dirk Passivan | 4.5 | Man | 4 |
| Lucas Jung | 1 | Man | 5 |
| Natalie Passivan | 4.5 | Woman | 6 |
| Patrick Dorner | 3.5 | Man | 7 |
| Svenja Erni | 3.5 | Woman | 8 |
| Walter Vlaanderen | 4.5 | Man | 9 |

# 3 Results

The results clearly identify compositions with the highest probability of being optimal. The probabilities of each player being included in the optimal line-up are also estimated (see Figure 1). Annabel Breuer, Correy Rossi, Dirk Passivan, and Walter Vlaanderen stand out. This demonstrates valuable information derived from the results.

Additionally, as we can observe in Figure 2, the methodology allows for calculating compatibility probabilities between players on the court. Looking at this figure, it becomes evident that Patrick Dorner is the player who best completes the group composed by {Annabel Breuer, Correy Rossi, Dirk Passivan, Walter Vlaanderen}.

Finally, it is also possible to estimate outcomes in the absence of a player due to suspension or injury, specifically considering line-ups without that player.

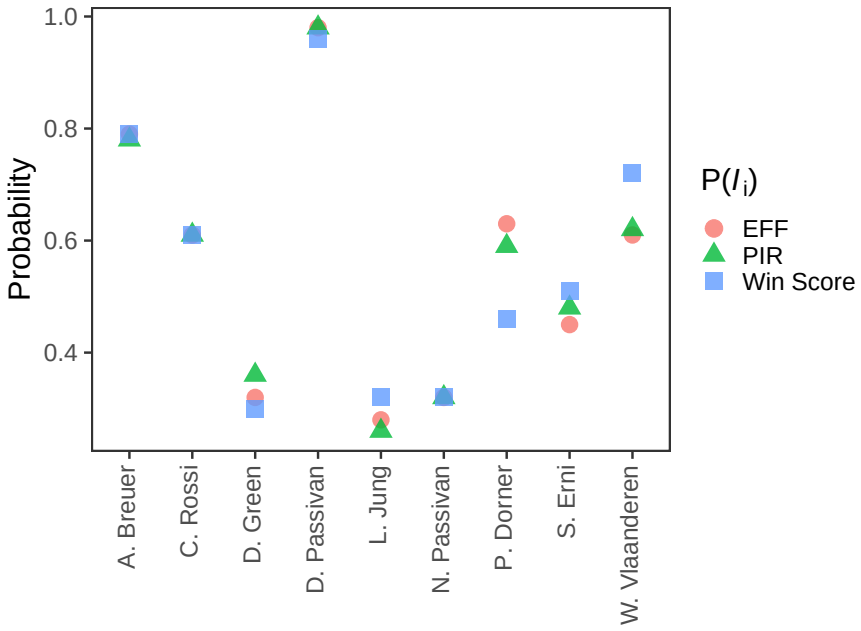



Figure 1: Estimated probabilities $P(I_i \mid \mathcal{D})$ of player $i$ being included in the optimal line-up team based on EFF, PIR and Win Score metrics.

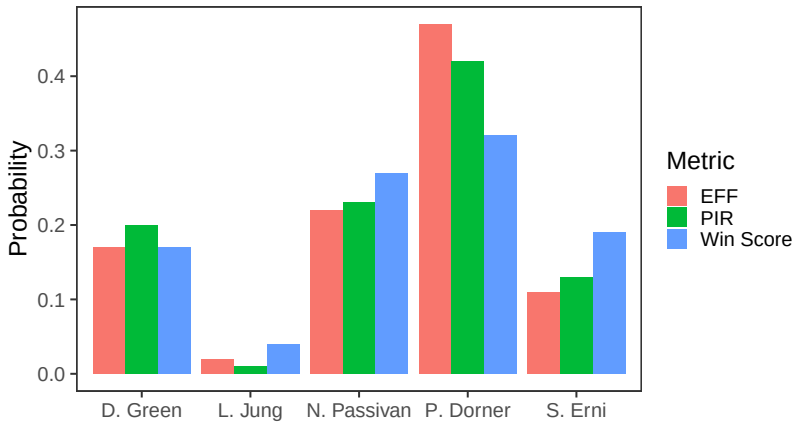



Figure 2: Estimated probabilities of the inclusion of players Dejon Green, Lucas Jung, Natalie Passivan, Patrick Dorner and Svenja Erni in the line-up team given that players Annabel Breuer, Correy Rossi, Dirk Passivan and Walter Vlaanderen have already been chosen, based on EFF, PIR and Win Score metrics.

# 4 Conclusions

This work focuses on combined metrics of positive and negative actions, but it can be easily complemented with other types of metrics. Our proposal is directly applicable to other sports for people with physical disabilities using the PCPS, such as powerchair football, powerchair hockey, or wheelchair rugby. It can even be adapted to sports without PCPS by replacing these constraints with specific limitations to each discipline. In conclusion, this methodology has great potential for use by coaches and technical staff, providing valuable insights into ideal compositions for upcoming games, evaluating both own and opposing players, and identifying strengths and weaknesses of different strategies.

# Acknowledgements

Gabriel Calvo's research was funded by the Spanish Ministry of Education and Professional Training, grant FPU18/03101. This paper is part of the project PID2022-136455NB-I00, funded by Ministerio de Ciencia, Innovación y Universidades of Spain (MCIN/AEI/10.13039/501100011033/FEDER, UE) and the European Regional Development Fund.

# References

Calvo, G., Armero, C., Grimm, B., and Ley, C. (2023). Selecting the best compositions of a wheelchair basketball team: a data-driven approach. *arXiv preprint arXiv:2310.03417*.

# A Multivariate Multilevel Longitudinal Functional Model for Repeatedly Observed Human Movement Data

Edward Gunning[1] & Steven Golovkine[1] & Andrew J. Simpkin[2] & Aoife Burke[3] & Sarah Dillon[4] & Shane Gore[3] & Kieran Moran[3] & Siobhan O'Connor[3] & Enda Whyte[3] & Norma Bargary[1]

[1] *MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland, firstname.lastname@ul.ie*

[2] *School of Mathematical and Statistical Sciences, University of Galway, Ireland, andrew.simpkin@universityofgalway.ie*

[3] *Centre for Injury Prevention and Performance, Athletic Therapy and Training; School of Health and Human Performance, Dublin City University, Dublin, Ireland, firstname.lastname@dcu.ie*

[4] *School of Allied Health, Faculty of Education and Health Science, University of Limerick, Limerick, Ireland, sarah.dillon@ul.ie*

**Résumé.** La recherche en biomécanique et en mouvement humain implique souvent la mesure régulière de plusieurs variables cinématiques ou cinétiques tout au long d'un mouvement, ce qui produit des données qui se présentent sous la forme de courbes, lisses, multivariées et variables dans le temps. Ces données se prêtent naturellement à l'analyse des données fonctionnelles. De plus, il est courant d'enregistrer le même mouvement de manière répétée pour chaque individu, ce qui permet d'obtenir des courbes corrélées pouvant être considérées comme des données fonctionnelles longitudinales.

Nous présentons une nouvelle approche de modélisation de données fonctionnelles longitudinales multivariées et hiérarchiques, en l'appliquant à des données cinématiques de coureurs amateurs recueillies lors d'une course sur tapis roulant. Pour chaque foulée, les angles de la hanche, des genoux et des chevilles des coureurs sont modélisés conjointement comme des fonctions multivariées dépendantes de covariables spécifiques au sujet. Des effets aléatoires fonctionnels multivariés variant longitudinalement sont utilisés pour capturer la dépendance entre les foulées adjacentes et les changements dans les fonctions multivariées au cours de la course. Nous représentons chaque observation en la décomposant dans une base en composantes principales fonctionnelles multivariées et nous modélisons les coefficients de base grâce des modèles scalaires longitudinaux à effets mixtes. Les effets aléatoires prédits sont utilisés pour comprendre et visualiser les changements dans les données fonctionnelles multivariées au cours de la course.

Dans notre application, cette méthode nous permet de quantifier les effets des covariables scalaires sur les données fonctionnelles multivariées. Il en résulte un effet statistiquement significatif de la vitesse de course au niveau des articulations de la hanche, du genou et de la cheville. L'analyse des effets aléatoires prédits révèle que la cinématique des individus est généralement stable, mais certains individus présentent de fortes variations au cours de la course.

**Mots-clés.** Analyse de données fonctionnelles longitudinales, Donnéees fonctionnelles multivariées, Analyse cinématique, Modèle mixte.

**Abstract.** Biomechanics and human movement research often involves measuring multiple kinematic or kinetic variables regularly throughout a movement, yielding data that present as smooth, multivariate, time-varying curves and are naturally amenable to functional data analysis. It is now increasingly common to record the same movement repeatedly for each individual, resulting in curves that are serially correlated and can be viewed as longitudinal functional data.

We present a new approach for modelling multivariate multilevel longitudinal functional data, with application to kinematic data from recreational runners collected during a treadmill run. For each stride, the runners' hip, knee and ankle angles are modelled jointly as smooth multivariate functions that depend on subject-specific covariates. Longitudinally varying multivariate functional random effects are used to capture the dependence among adjacent strides and changes in the multivariate functions over the course of the treadmill run. We represent each observation using a multivariate functional principal components basis and model the basis coefficients using scalar longitudinal mixed effects models. The predicted random effects are used to understand and visualise changes in the multivariate functional data over the course of the treadmill run.

In our application, the method quantifies the effects of scalar covariates on the multivariate functional data, revealing a statistically significant effect of running speed at the hip, knee and ankle joints. Analysis of the predicted random effects reveals that individuals' kinematics are generally stable but certain individuals who exhibit strong changes during the run can also be identified.

**Keywords.** Longitudinal functional data analysis, Multivariate functional data, Kinematic analysis, Mixed-effects model

# 1 Introduction

Longitudinal functional data analysis (LFDA) concerns the modelling of the dependence among functions due to correlation over a longer (or different) timescale than the one on which they are measured. Examples include daily activity functions measured consecutively for a number of days for several subjects (Goldsmith *et al.*, 2015) or brain imaging profiles of patients measured at several hospital visits (Greven *et al.*, 2010), see Park and Staicu (2015).

Our motivating dataset comes from the Dublin City University running injury surveillance (RISC) study, where kinematic data from recreational runners were captured during a treadmill run with the goal of understanding running technique and its link to injury. We focus on modelling the sagittal plane hip, knee and ankle angles because the majority of running-related injuries occur in the lower limbs. During the treadmill run, the kinematic data were recorded for a large number of consecutive strides for each individual (see Figure 1). They were then segmented into individual strides, as a single stride is considered the

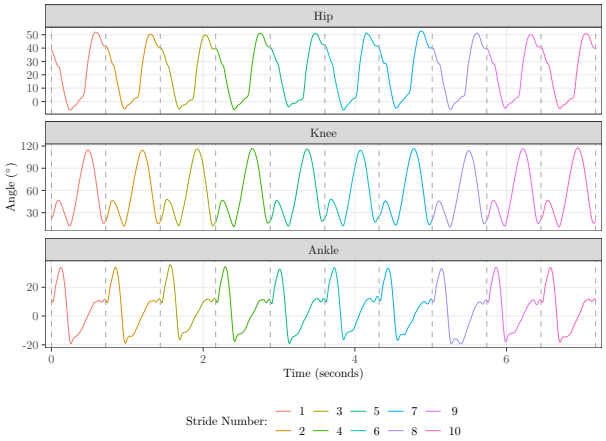



Figure 1: The right sagittal hip, knee and ankle angles of a single participant in the RISC dataset for the first ten strides of their treadmill run. The dashed vertical lines indicate touch down (i.e., when the foot first touches the ground), which represents the start and end of each stride.

most basic unit of analysis.

In this work, we want to employ a multivariate approach to capture the dependence among multiple joints (i.e., the hip, knee and ankle angles), rather than performing separate univariate analyses for each location. From an applied perspective, understanding the dependence (or co-ordination) among multiple joints is crucial for fully describing movement patterns (Glazier, 2021). Moreover, the participants were measured on both sides of the body, which adds a hierarchical structure to the data. We also need to include scalar covariate information in our model, e.g., sex, running speed and injury status. This motivates the development of a *multivariate multilevel longitudinal functional model*. The dataset contains more than 40000 multivariate functional observations from 284 unique individuals. To the best of our knowledge, this is the first piece of work to develop statistical methodology to appropriately analyse repeatedly observed multivariate kinematic data in human movement biomechanics.

# 2 Model

We denote the multivariate functional observation from the $l$th stride for the $i$th individual on side $j$ as

$$\mathbf{y}_{ijl}(t) = \left(y_{ijl}^{(hip)}(t), y_{ijl}^{(knee)}(t), y_{ijl}^{(ankle)}(t)\right)^\top, l = 1, \dots n_{ij}, j \in \{\text{left, right}\} \text{ and } i = 1, \dots, N,$$

where $N$ is the total number of individuals, $n_{ij}$ is the total number of strides taken by individual $i$ on side $j$, and $t \in [0, 100]$ is a normalised *functional time* interval with 0 representing the start of a stride and 100(%) representing the end. We also introduce a (normalised) *longitudinal time* variable $T \in [0, 1]$, such that $T_{ijl}$ indexes the time in the treadmill run at which

stride $l$ occurs on side $j$ for subject $i$. The start of the treadmill run is at $T = 0$ and the end at $T = 1$. Finally, we let $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijA})^\top$ denote the scalar covariates for subject $i$ on side $j$. The covariates could be subject specific (e.g., sex, height) or subject-and-side specific (e.g., an indicator for a subject's dominant side). We assume that the covariates are fixed across strides and hence $\mathbf{x}_{ij}$ is not indexed by $l$.

Our proposed multivariate multilevel longitudinal functional model is

$$\mathbf{y}_{ijl}(t) = \boldsymbol{\beta}_0(t, T_{ijl}) + \sum_{a=1}^{A} x_{ija}\boldsymbol{\beta}_a(t) + \mathbf{u}_i(t, T_{ijl}) + \mathbf{v}_{ij}(t, T_{ijl}) + \boldsymbol{\varepsilon}_{ijl}(t).$$

where $\boldsymbol{\beta}_0(t, T_{ijl})$, $\mathbf{u}_i(t, T_{ijl})$ and $\mathbf{v}_{ij}(t, T_{ijl})$ are the multivariate intercept function, the subject-specific and the subject and side-specific multivariate functional random intercepts that varies smoothly in both functional and longitudinal time, $\boldsymbol{\beta}_a(t)$ is the multivariate functional fixed effect corresponding to the $a$th scalar covariate and $\boldsymbol{\varepsilon}_{ijl}(t)$ is the smooth multivariate functional random error that is specific to observation $\mathbf{y}_{ijl}(t)$. We assume that $\mathbf{y}_{ijl}(t)$ are centered.

The intercept function $\boldsymbol{\beta}_0(t, T)$ is assumed to be a smooth bivariate function of both functional time $t$ and longitudinal time $T$. For $a = 1, \ldots, A$, the fixed effect $\boldsymbol{\beta}_a(t)$ captures the influence of the $a$th scalar covariate on the expected level and shape of the response (Bauer *et al.*, 2018). We assume that the fixed effects are constant across $T$, which implies that the scalar covariates affect the average running kinematics, rather than the kinematics at any particular point in the treadmill run. For $i = 1, \ldots, N$, the subject-specific random intercept $\mathbf{u}_i(t, T)$ captures correlation among observations from the same subject and the subject-and-side-specific random intercepts $\mathbf{v}_{ij}(t, T)$ capture correlation among observations from the same subject and side. These functions are assumed to be independent realisations of mean-zero multivariate Gaussian processes with matrix-valued covariance function $\mathbf{Q}(t, t', T, T')$ and $\mathbf{R}(t, t', T, T')$. Finally, the random errors are assumed to be independent realisations of a zero-mean multivariate Gaussian process with matrix-valued covariance function $\mathbf{S}(t, t')$. The multivariate functional random error represents the deviation that is specific to observation $\mathbf{y}_{ijl}(t)$. It is further assumed that the processes $\mathbf{u}_i(t, T)$, $\mathbf{v}_{ij}(t, T)$ and $\boldsymbol{\varepsilon}_{ijl}(t)$ are mutually uncorrelated.

## 2.1 Basis Representation of the Multivariate Functions

We first represent each multivariate functional observation by a basis expansion

$$\mathbf{y}_{ijl}(t) = \sum_{k=1}^{K} y^{\star}_{ijl,k}\boldsymbol{\psi}_k(t).$$

The basis functions $\{\boldsymbol{\psi}_k(t)\}_{k=1}^K$ are multivariate functions and $y^{\star}_{ijl,k}$ are scalar basis coefficients that weight the basis functions to produce the functional observations. The functions $\{\boldsymbol{\psi}_k(t)\}_{k=1}^K$ are estimated using a multivariate functional principal components analysis (MFPCA, Happ and Greven, 2018). We choose $K$, the number of functions to retain, such that a high percentage (e.g., 99.5%) of the variance in the data is explained. This allows the basis coefficients to be treated as transformed data rather than estimated parameters and modelled in place of the observed multivariate functions (Morris *et al.*, 2011).

## 2.2 Modelling the Basis Coefficients

We model the matrix $\mathbf{Y}^\star$ of basis coefficients in place of the observed multivariate functional data. We make the simplifying assumption that each of the $K$ basis coefficients (i.e., each column of $\mathbf{Y}^\star$) can be modelled separately (Morris and Carroll, 2006). The model for the $k$th basis coefficient is

$$y^\star_{ijl,k} = \beta^\star_{0,k}(T_{ijl}) + \sum_{a=1}^{A} x_{ia}\beta^\star_{a,k} + u^\star_{i,k}(T_{ijl}) + v^\star_{ij,k}(T_{ijl}) + \varepsilon^\star_{ijl,k}, \tag{1}$$

which is a multilevel functional model in longitudinal time $T$ (Di *et al.*, 2009). We choose to parameterise the longitudinally varying functions using a small number of unpenalised basis functions, because changes are expected to be smooth and simple. We also use the same set of basis functions $\{\xi_d(T)\}_{d=1}^{D}$ to represent each longitudinally varying term, giving

$$\beta^\star_{0,k}(T) = \sum_{d=1}^{D} \beta^\star_{0,k,d}\,\xi_d(T), \quad u^\star_{i,k}(T) = \sum_{d=1}^{D} u^\star_{i,k,d}\,\xi_d(T) \quad \text{and} \quad v^\star_{ij,k}(T) = \sum_{d=1}^{D} v^\star_{ij,k,d}\,\xi_d(T).$$

We use a small number of natural cubic B-spline basis functions to represent each term. Substituting the basis function evaluations into model (1) gives, for the $k$th basis coefficient, the model

$$y^\star_{ijl,k} = \sum_{d=1}^{D} \beta^\star_{0,k,d}\,\xi_d(T_{ijl}) + \sum_{a=1}^{A} x_{ia}\beta^\star_{a,k} + \sum_{d=1}^{D} u^\star_{i,k,d}\,\xi_d(T_{ijl}) + \sum_{d=1}^{D} v^\star_{ij,k,d}\,\xi_d(T_{ijl}) + \varepsilon^\star_{ijl,k},$$

where $(u^*_{i,k,1}, \ldots, u^*_{i,k,D})^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^*_k)$, $(v^*_{ij,k,1}, \ldots, v^*_{ij,k,D})^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^*_k)$ and $\varepsilon^*_{ijl,k} \sim \mathcal{N}(0, s_k)$. This is a scalar linear mixed effects model. The matrices $\mathbf{Q}^*_k$ and $\mathbf{R}^*_k$ are of dimension $D \times D$ and contain $D(D+1)/2$ free parameters to estimate.

## 2.3 Reconstructing the Model Terms

### 2.3.1 Fixed Effects

Rather than inspect individual parameter estimates, it is more natural to combine the estimated parameters across the basis coefficients to reconstruct and estimate the functional model terms. The estimated intercept function is given by

$$\widehat{\boldsymbol{\beta}}_0(t, T) = \sum_{k=1}^{K}\sum_{d=1}^{D} \widehat{\beta}^\star_{0,k,d}\,\xi_d(T)\boldsymbol{\psi}_k(t),$$

where $\widehat{\beta}^\star_{0,k,d}$ denotes the estimate of $\beta^\star_{0,k,d}$ from the mixed effects model. Likewise, the estimate of the functional fixed effect of the $a$th scalar covariate is given by

$$\widehat{\boldsymbol{\beta}}_a(t) = \sum_{k=1}^{K} \widehat{\beta}^\star_{a,k}\boldsymbol{\psi}_k(t), \quad a = 1 \ldots, A.$$

The estimates of $\widehat{\mathrm{Var}}(\widehat{\beta}^{\star}_{a,k})$ from the mixed effects model can be combined across $k$ to construct approximate pointwise and simultaneous confidence bands for $\boldsymbol{\beta}_a(t)$, as described in (Gunning *et al.*, 2023).

### 2.3.2 Covariance Structures

The matrix-valued covariance functions $\mathbf{Q}(t,t',T,T')$ and $\mathbf{R}(t,t',T,T')$ implied by the model are

$$\mathbf{Q}(t,t',T,T') = \mathbb{E}[\mathbf{u}_i(t,T)\ \mathbf{u}_i(t',T')^\top] = \boldsymbol{\Psi}(t)^\top(\mathbb{I}_K \otimes \boldsymbol{\xi}(T))^\top \mathbf{Q}^*(\mathbb{I}_K \otimes \boldsymbol{\xi}(T'))\boldsymbol{\Psi}(t'),$$
$$\mathbf{R}(t,t',T,T') = \mathbb{E}[\mathbf{v}_{ij}(t,T)\ \mathbf{v}_{ij}(t',T')^\top] = \boldsymbol{\Psi}(t)^\top(\mathbb{I}_K \otimes \boldsymbol{\xi}(T))^\top \mathbf{R}^*(\mathbb{I}_K \otimes \boldsymbol{\xi}(T'))\boldsymbol{\Psi}(t'),$$

where $\boldsymbol{\Psi}(t)$ is the $K \times 3$ matrix containing the basis functions, $\boldsymbol{\xi}(T) = (\xi_1(T), \ldots, \xi_D(T))^\top$, $\mathbf{Q}^*$ (resp. $\mathbf{R}^*$) is the block-diagonal matrix containing the matrices $\mathbf{Q}^*_1, \ldots, \mathbf{Q}^*_K$ (resp. $\mathbf{R}^*_1, \ldots, \mathbf{R}^*_K$) along its diagonal. Finally, the within-function covariance is

$$\mathbf{S}(t,t') = \boldsymbol{\Psi}(t)^\top \mathbf{S}^* \boldsymbol{\Psi}(t'), \quad \mathbf{S}^* = \mathrm{diag}\{s_1, \ldots, s_K\}.$$

### 2.3.3 Individual Trajectories

Our methodology facilitates the prediction of subject-specific and subject-and-side-specific trajectories at any point in the treadmill run. The prediction of the subject-specific multivariate functional random intercept at any $T \in [0,1]$ is given by

$$\widehat{\mathbf{u}}_i(t,T) = \sum_{k=1}^{K}\sum_{d=1}^{D} \widehat{u}^{\star}_{i,k,d}\ \xi_d(T)\boldsymbol{\psi}_k(t), \quad i = 1, \ldots, N,$$

where $\widehat{u}^{\star}_{i,k,d}$ is the Best Linear Unbiased Predictor (BLUP) of $u^{\star}_{i,k,d}$ from the linear mixed effects model. The subject-and-side specific deviation is obtained analogously as

$$\widehat{\mathbf{u}}_i(t,T) + \widehat{\mathbf{v}}_{ij}(t,T) = \sum_{k=1}^{K}\sum_{d=1}^{D} \big(\widehat{u}^{\star}_{i,k,d} + \widehat{v}^{\star}_{ij,k,d}\big)\xi_d(T)\boldsymbol{\psi}_k(t),\ i = 1, \ldots, N,\ \text{and } j \in \{\text{left}, \text{right}\}.$$

The predicted trajectories can be used, for example, to investigate change in technique over the course of the treadmill run as measured by the rate of change with respect to $T$.

# 3 Application

## 3.1 Data Collection, Extraction and Preparation

Recreational runners aged between 18 and 64 years of age with no history of injury in the last three months were recruited as participants for the RISC study. Prior to the baseline testing session, in which the kinematic data were collected, the participants completed an

online survey regarding their injury history, training history and demographics. They ran for three minutes at a self-selected speed that represented their typical training pace, while kinematic data were collected using a 17-camera, three-dimensional motion analysis system for the first full minute of the run. The motion data (i.e., marker trajectories) were sampled at a rate of 200Hz and filtered using a fourth-order zero-lag Butterworth filter at 15 Hz to smooth out observational errors. From the filtered trajectories, the sagittal plane hip, knee and ankle angles were extracted bilaterally for the first minute of the treadmill run.

The long sequences of kinematic measurements (e.g., Figure 1) were segmented into individual strides based on the initial contact of the foot with the ground, which was identified using a custom algorithm. The univariate functional data for each stride were time normalised and registered to the point of the maximum knee flexion angle, which is a clear and easily identifiable landmark in each stride. Within each dimension, 80 cubic B-spline basis functions were used to provide a near-lossless representation of the univariate functions. For each stride, the longitudinal time variable $T$ was created based on the time at which that stride started, with $T = 0$ representing the start of the subject's capture period. This variable was normalised by dividing by the subject's maximum capture time, so that $T \in [0, 1]$. The MFPCA, computed from the univariate basis expansions, yielded $K = 27$ basis functions to explain 99.5% of the variance in the multivariate functional data.

## 3.2 Modelling Results

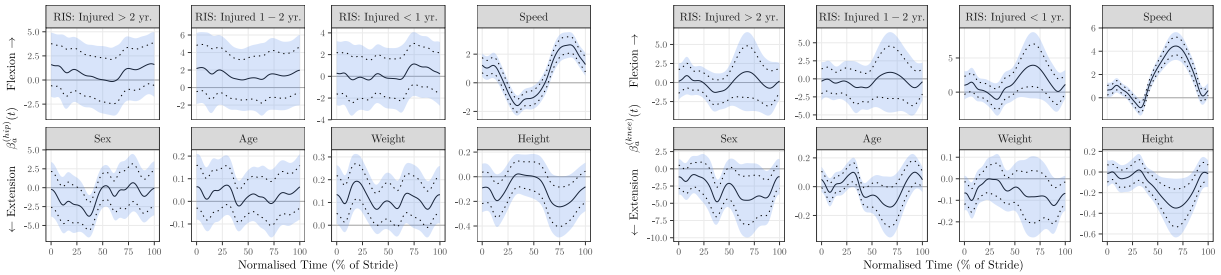
A constant function and four natural cubic B-splines were used as longitudinal basis functions, with unstructured $\mathbf{Q}_k^*$ and $\mathbf{R}_k^*$ matrices. We consider the following model:

$$\begin{aligned}\mathbf{y}_{ijl}(t) = \boldsymbol{\beta}_0(t, T_{ijl}) + \sum_{a=1}^{3} x_{ia}\boldsymbol{\beta}_a(t) + \ \text{speed}_i \times \boldsymbol{\beta}_4(t) + \text{sex}_i \times \boldsymbol{\beta}_5(t) + \text{age}_i \times \boldsymbol{\beta}_6(t) \\ + \text{weight}_i \times \boldsymbol{\beta}_7(t) + \text{height}_i \times \boldsymbol{\beta}_8(t) + \mathbf{u}_i(t, T_{ijl}) + \mathbf{v}_{ij}(t) + \boldsymbol{\varepsilon}_{ijl}(t),\end{aligned}$$

where $x_{i1}, x_{i2}$ and $x_{i3}$ are dummy-coded variables representing the "Injured more than 2 years ago", "Injured 1-2 years ago" and "Injured less than 1 year ago" categories of the retrospective injury status variable, where the reference category is "Never injured", $\text{speed}_i$ is the self-selected running speed of subject $i$ in $\text{km}\,\text{h}^{-1}$, $\text{sex}_i$ is a dummy-coded variable for the sex of subject $i$ (0 = male, 1 = female), $\text{age}_i$ is the age of subject $i$ in years, $\text{weight}_i$ is the weight of subject $i$ in kilograms and $\text{height}_i$ is the height of subject $i$ in centimetres. All numeric variables were centered to make the intercept function more interpretable.

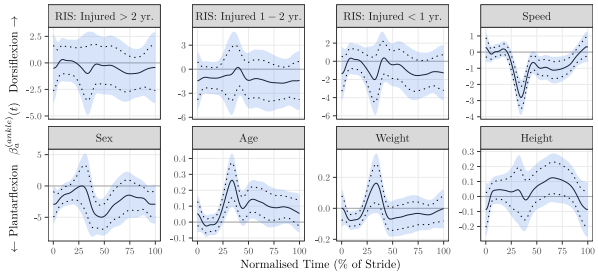
### 3.2.1 Fixed Effects

Analysis of the functional coefficients of the longitudinal basis functions used to model the intercept revealed that it was approximately constant in the longitudinal direction. Figure 2 displays the estimated coefficient functions that capture the effects of scalar covariates in our model. In all three dimensions, the simultaneous confidence bands for the retrospective injury status coefficient functions contain zero (solid grey horizontal line) for all $t$, indicating that




(a) Hip

(b) Knee



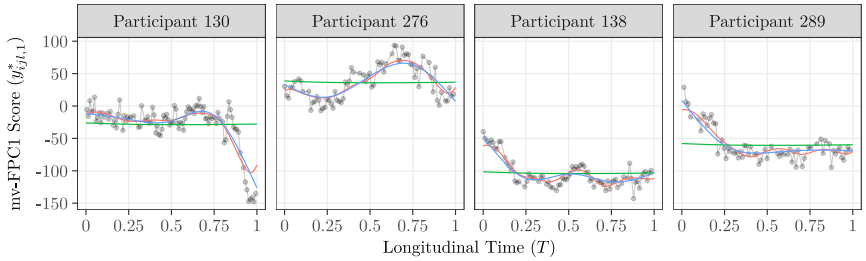

(c) Ankle

Figure 2: The estimated coefficient functions of the fixed effects from the fitted model. The black solid line represents the point estimate, the dotted black lines indicate pointwise 95% confidence intervals and the light blue ribbons represent 95% simultaneous confidence bands.

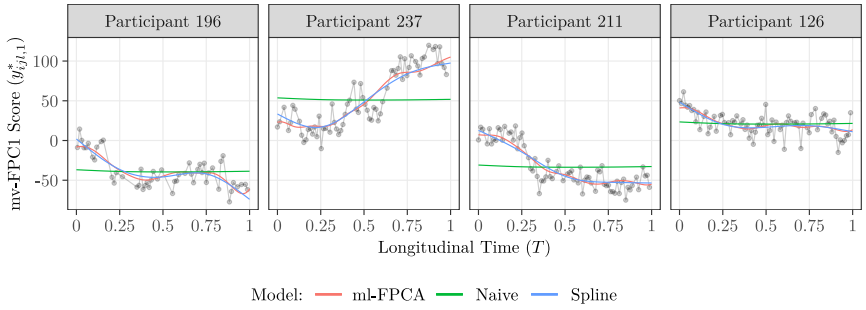
there is no evidence of a significant difference between any of the categories and the reference category of "Never injured". We observe a strong, noticeable effect of self-selected running speed in all three dimensions, as the simultaneous confidence bands only contain zero around the time that the point estimate crosses 0. Running at a higher speed is associated with greater hip flexion at initial contact and late in the swing phase ($t > 60\%$) and greater hip extension around the time of toe-off ($t \approx 38\%$), greater knee flexion which is most pronounced in the stance phase around the time of peak knee flexion angle ($t \approx 69\%$) and increased ankle plantarflexion which is most pronounced around the time of maximum plantarflexion ($t \approx 38\%$). The coefficient functions for the effect of sex are large in magnitude, reaching almost 5° in the knee and ankle. However, the corresponding confidence bands are wide and contain zero for almost all $t$, indicating a lot of uncertainty about this effect. There is limited evidence of an age, height or weight effect. Although the simultaneous confidence bands for these coefficient functions do not contain zero at certain points, the magnitude of each effect is small.

### 3.2.2 Random Effects

We present analysis of the fitted subject-and-side specific trajectories, which are obtained as BLUPs of the random effects. Figure 3 displays fits for subjects that were chosen according to summaries from the model. Firstly, we calculated the integrated squared first derivative with




(a) Rate of change




(b) Overall change

Figure 3: Observed and fitted values of the coefficients for subjects identified based on summaries from the model. **(a)** The top four subjects based on the integrated squared first derivative of their fitted trajectory with respect to longitudinal time. **(b)** The top four subjects based on the overall change during the treadmill run.

respect to longitudinal time of each subject's fitted profile, which provides a measure of the rate of change (or deviation from a constant fit) over the course of the treadmill run. Figure 3 (a) displays the first coefficients for the top four subjects ranked according to this metric. For ease of interpretation, we have only displayed the left side observations. All four subjects exhibit non-stationary patterns that are captured well by the longitudinal models. The naive model, which assumes that each individual's deviation is constant across longitudinal time, is inadequate. Figure 3 (b) displays another four subjects ranked according to the overall change in the subject's fitted profile over the course of the run, calculated as the absolute difference between the subjects' fitted profiles at $T = 0$ and $T = 1$. Non-stationary trends, which cannot be captured by the naive model, are evident again. It should be noted that these summaries were computed based on the full multivariate function but we have displayed the first coefficient. However, this coefficient captured the largest amount of variance in the longitudinal direction, so it is a reasonable choice.

As the coefficients in Figure 3 are a level of abstraction away from the multivariate functional data, we examine the fitted multivariate functions for a single individual. Based on Figure 3 (b), we choose to display Participant 237 because they exhibited a consistent, almost-linear evolution. Figure 4 (c) and (d) display the motion-capture animation at the time of peak knee flexion angle for this subject at the start (stride 1) and end (stride 80) of the treadmill run, respectively. The difference in the two pictures reflects the changes across longitudinal time, in particular the greater knee flexion at the end of the treadmill run.

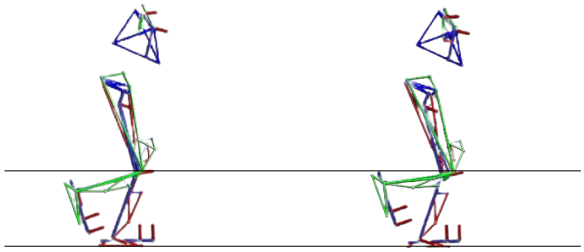


Figure 4: Individual analysis for Participant 237. **(Left)** The motion-capture animation for this subject at the time of peak knee flexion angle at the start of the treadmill run. **(Right)** The motion-capture animation for this subject at the time of peak knee flexion angle at the end of the treadmill run.

# Bibliography


Bauer, A., Scheipl, F., Küchenhoff, H., & Gabriel, A.-A. (2018), An introduction to semi-parametric function-on-scalar regression. *Statistical Modelling*, 18(3-4), pp. 346-364.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009), Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1), pp.458-488.

Happ, C., & Greven, S. (2018), Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113(522), pp. 649-659.

Glazier, P. S. (2021), Beyond animated skeletons: How can biomechanical feedback be used to enhance sports performance? *Journal of Biomechanics*, 129, 110686.

Goldsmith, J., Zipunnikov, V., & Schrack, J. (2015), Generalized multilevel function-on-scalar regression and principal component analysis *Biometrics*, 71(2), pp. 344-353.

Greven, S., Crainiceanu, C. M., Caffo, B., & Reich, D. (2010), Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, 4, pp. 1022-1054.

Gunning, E., Golovkine, S., Simpkin, A. J., Burke, A., Dillon, S., Gore, S., Moran, K. A., O'Connor, S., Whyte, & Bargary, N. (2023). Analyzing Kinematic Data from Recreational Runners using Functional Data Analysis. *Under Review*.

Morris, J. S., & Carroll, R. J. (2006), Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(2), pp. 179-199.

Morris, J. S., Baladandayuthapani, V., Herrick, R. C., Sanna, P., & Gutstein, H. (2011), Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *The Annals of Applied Statistics*, 5(2A), pp. 894-923.

Park, S. Y., & Staicu, A.-M. (2015), Longitudinal functional data analysis. *Stat*, 4(1), 212–226.

# Modèles de Markov dérivants pour l'apprentissage de l'escalade

Emmanouil-Nektarios Kalligeris [1,2,5] & Vlad Stefan Barbu [2,3] & Guillaume Hacques [1] & Ludovic Seifert[1,4] & Nicolas Vergne [2]

[1] *Centre d'études des Transformations des Activités Physiques et Sportives, Université de Rouen Normandie, Boulevard Siegfried, 76821 Mont-Saint-Aignan, Rouen, France, guillaume.hacques@univ-rouen.fr*

[2] *Laboratoire de Mathématique Raphaël Salem, Université de Rouen Normandie, Avenue de l'Université, BP.12, 76801 Saint-Étienne du Rouvray, Rouen, France, nicolas.vergne@univ-rouen.fr*

[3] *Centre for Demographic Research "Vladimir Trebici", "Costin C. Kiritescu" National Institute of Economic Research of Romanian Academy, Romania, barbu@univ-rouen.fr*

[4] *Institut Universitaire de France (IUF), Paris, France, ludovic.seifert@univ-rouen.fr*

[5] *University of Sheffield, School of Mathematics and Statistics, UK, e.kalligeris@sheffield.ac.uk*

**Résumé.**

Cette recherche explore la dynamique de l'apprentissage à long terme de l'escalade en utilisant des modèles de Markov dérivants. L'escalade implique des prises de décisions complexes qui exigent une coordination visuo-motrice efficace et une exploration attentive de l'environnement. Les modèles de Markov dérivants représentent une catégorie de processus de Markov contraints hétérogènes adaptés à la modélisation de données présentant une certaine hétérogénéité. En appliquant ces modèles aux données réelles de compétences visuo-motrices, notre objectif est de dévoiler la dynamique persistante de l'apprentissage de l'escalade. Pour ce faire, nous conduisons une étude de cas réelle, fournissant des résultats qui (i) contribuent à la compréhension de l'acquisition de compétences dans des environnements physiquement contraints et (ii) offrent un aperçu du rôle crucial de l'exploration et de la coordination visuo-motrice dans le processus d'apprentissage.
**Mots-clés.** Modèles de Markov dérivants, Statistiques pour le sport, Dynamique persistante, Coordination visuo-motrice, Escalade.

**Abstract.** This research delves into the dynamics of long-term learning in climbing, employing Drifting Markov Models. Climbing represents a complex decision-making challenge, necessitating effective coordination between visual and motor skills and exploration of the surroundings. Drifting Markov Models constitute a category of constrained, heterogeneous Markov processes designed for modeling data showcasing heterogeneity. Through the application of these models to real-world visual-motor skill data, our goal is to unveil the enduring dynamics of learning in climbing. To achieve this, a genuine case study is conducted through an experiment, yielding results that (i) contribute to comprehending skill acquisition in physically demanding settings and (ii) offer insights into the pivotal role of exploration and visual-motor coordination in the learning process.
**Keywords.** Drifting Markov Models, Sport statistics, Persistent Dynamics, Visual- Motor Coordination, Climbing.

# 1 Introduction

La modélisation de la dynamique d'apprentissage à long terme est depuis longtemps un sujet d'intérêt dans le domaine de la psychologie (Adams, 1961; Hallett & Grafman, 1997; Irion, 1948). Dans le contexte de la pratique et de l'apprentissage, l'exploration est également la découverte des différentes possibilités d'atteindre l'objectif d'une tâche. Elle reflète donc la navigation de l'individu dans l'espace de travail visuo-moteur pour obtenir une solution efficace et efficiente à la tâche (Newell et al., 1989; Komar, Seifert, Vergne, & Newell, 2023).

L'escalade constitue une tâche complexe de prises de décisions, où les données sur les compétences visuo-motrices peuvent s'avérer essentielles pour comprendre la dynamique d'apprentissage des grimpeurs. Il ne fait aucun doute que les grimpeurs doivent s'appuyer sur de multiples sources d'information (par exemple, visuelles, auditives, tactiles) pour analyser leur environnement et les aider dans leurs actions. Par exemple, Hacques, Komar et Seifert (2021) ont montré que l'activité exploratoire évolue avec la pratique d'une tâche d'escalade. Ils ont conclu que, pendant plusieurs sessions, les individus diminuaient l'utilisation des mouvements exploratoires de la main pour s'appuyer davantage sur les informations visuelles afin de guider leurs actions de manière efficace. Comme on peut s'en douter, la coordination efficace des mouvements des apprenants dans des environnements physiquement contraints est un défi majeur pour la modélisation d'une tâche décisionnelle complexe telle que l'escalade.

L'hétérogénéité accompagnant presque toutes les tâches complexes, il est raisonnable de supposer que les compétences visuo-motrices sont régies par cette hétérogénéité. L'estimation de processus hétérogènes généraux étant difficile à mettre en oeuvre dans la pratique, Vergne (2008) a introduit la classe des modèles de Markov dérivants (DMM pour *Drifting Markov Models*). Les DMM reposent sur l'idée simple qu'au lieu d'adapter une matrice de transition unique à l'ensemble de la séquence d'observations, la matrice de transition est autorisée à dériver le long de la séquence. Ainsi, à chaque position, une matrice de transition différente est obtenue, avec comme contrainte une évolution douce en fonction de la position. En d'autres termes, les DMM constituent une classe particulière de processus de Markov hétérogènes contraints qui peuvent s'avérer flexibles pour des séquences telles que les données d'aptitudes visuo- motrices. Les DMM se sont récemment révélés être une approche prometteuse pour l'apprentissage d'une coordination efficace entre les membres dans la brasse (Komar et al., 2023).

Dans ce travail, nous étudions la dynamique d'apprentissage du grimpeur sur plusieurs sessions, c'est-à-dire sur une longue échelle de temps, à travers le mécanisme des DMMs. Plus précisément, la section 2 définit les DMMs et décrit leur estimation. La section 3 est consacrée à une étude de cas réel. Après avoir détaillé les caractéristiques de la collecte de données ainsi que le cadre expérimental, une méthode de clustering non supervisée est appliquée afin d'acquérir les groupes constituant les états du DMM. Enfin, la section 4 résume les résultats obtenus.

# 2 Modèles de Markov Dérivants

Les DMM sont une classe particulière de processus de Markov hétérogènes qui permettent à la matrice de transition de dériver (varier) le long de la séquence des observations.
Considérons un système aléatoire avec un espace d'états fini $A = \{1, 2, \dots, s\}$ et soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité. De plus, soient $(\Pi_0(u,v))_{u,v\in A}$ et $(\Pi_1(u,v))_{u,v\in A}$, des matrices de transition d'un modèle de Markov d'ordre 1 sur l'espace d'états $A$.

## 2.1 Définition d'un modèle de Markov dérivant

**Definition 1 (Chaîne de Markov dérivante d'ordre 1 et de degré 1)** *Une suite de variables aléatoires $X_0, \dots, X_n$ dans un espace d'état $A$ est appelée une chaîne de Markov dérivante d'ordre 1, de degré 1 et de longueur $n$ entre $\Pi_0(u,v)$ et $\Pi_1(u,v)$, si la distribution des $X_t$, $t = 1, \dots, n$, est définie par :*

$$\Pi_{\frac{t}{n}}(u,v) = P\left(X_t = v \middle| X_{t-1} = u\right) = \left(1 - \frac{t}{n}\right)\Pi_0(u,v) + \frac{t}{n}\Pi_1(u,v).$$

**Definition 2 (Chaîne de Markov dérivante d'ordre $k$ et de degré $d$))** *Soit $\Pi_{\frac{i}{d}}$, $i = 0, \dots, d$, des matrices stochastiques sur $A$. Une suite de variables aléatoires $X_0, \dots, X_n$ dans un espace d'état $A$ est appelée une chaîne de Markov dérivante d'ordre $k$, de degré $d$ et de longueur $n$ entre $\Pi_{\frac{i}{d}}(u_1, \dots, u_k, v)$, $i = 0, \dots, d,$, si la distribution des $X_t$, $t = 1, \dots, n$, est définie par :*

$$\Pi_{\frac{t}{n}}(u_1, \dots, u_k, v) = P\left(X_t = v \middle| X_{t-1} = u_k, \dots, X_{t-k} = u_1\right) = \sum_{i=0}^{d} p_i(t)\Pi_{\frac{i}{d}}(u_1, \dots, u_k, v).$$

*où $(u_1, \dots, u_k) \in A^{k+1}$ et $p_i(t), i = \{0, 1, 2, \dots\}$ sont les polynômes de Lagrange de degré $d$ tels que :*

$$\forall (i,j) \in \{0, \dots, d\}^2, p_i\left(\frac{nj}{d}\right) = \mathbb{I}_{\{i=j\}}.$$

Avec les notations de la définition précédente, on a : (1) pour $t = \frac{ni}{d}$, on a $\Pi_{\frac{t}{n}} = \Pi_{\frac{i}{d}}$, $i = 0, \dots, d$, et (2) $\forall t \in [0, n]$, $\sum_{v\in A} \Pi_{\frac{t}{n}}(u,v) = 1$.

Puisque les $p_i(t)$ sont les polynômes de Lagrange, choisis pour avoir des matrices $\Pi_{\frac{t}{n}}$ stochastiques $\forall t \in [0, n]$, ils sont obtenus facilement pour un DMM d'ordre 1 et de degré 1. En posant $t = 0$ et $t = 1$ on a :

- $\Pi_{\frac{t}{n}} = \Pi_0$, c-à-d, $p_0(0) = 1$ et $p_1(0) = 0$;
- $\Pi_{\frac{t}{n}} = \Pi_1$, c-à-d, $p_0(1) = 0$ et $p_1(1) = 1$;

ainsi (1) est vérifiée.

Il est clair que, lorsque le degré augmente, le DMM correspondant admet une expression plus compliquée.

## 2.2 Estimation d'un modèle de Markov dérivant

Pour estimer $\Pi_{\frac{i}{d}}(u,v)$, $i=0,\ldots,d$, on utilise une méthode point par point, que nous allons illustrer brièvement sur un DMM d'ordre $k$ et de degré $d$. La fonction à minimiser est la suivante :

$$\sum_{t=1}^{n}\sum_{u\in A^k}\sum_{v\in A}\mathbb{I}_{\{X_{t-k}\ldots X_{t-1}=u\}}\left(\Pi_{\frac{t}{n}}(u,v)-\mathbb{I}_{\{X_t=v\}}\right)^2.$$

Pour déterminer les estimateurs $\widehat{\Pi}_{\frac{i}{d}}(u,v)$ de $\Pi_{\frac{i}{d}}(u,v)$, $\forall(u,v)\in A^k\times A$, nous devons résoudre le système suivant :

$$\left(\sum_{t=k}^{n}\mathbb{I}_{\{X_{t-k}\ldots X_{t-1}=u\}}p_i(t)p_j(t)\right)_{i,j\in[1,d]}\left(\widehat{\Pi}_{\frac{i}{d}}(u,v)\right)_{i\in[1,d]}=\left(\sum_{t=k}^{n}p_i(t)\mathbb{I}_{\{X_t=v,X_{t-k}\ldots X_{t-1}=u\}}\right)_{i\in[1,d]}.$$

Cette méthode d'estimation point par point, basée sur toutes les observations de la séquence, s'est avérée la plus efficace, en termes de vraisemblance. Pour plus d'informations sur cette méthode d'estimation ainsi que sur les propriétés des estimateurs obtenus, se référer à Vergne (2008) ; Barbu et Vergne(2019).

# 3 Matériel et Méthode

## 3.1 Participants, Mesures

L'expérience porte sur un total de 11 individus qui ont été enregistrés au cours de 10 sessions d'escalade. La première et la dernière session consistent en 6 essais chacune, tandis que les sessions intermédiaires consistent en 9 essais chacune. Il convient de noter que le niveau de compétence en escalade de chaque participant se situait dans le groupe inférieur selon l'échelle de l'*International Rock Climbing Research Association* (Draper et al., 2016). Le protocole est conforme à la Déclaration d'Helsinki et a été validé par le comité national d'éthique (ID : ANR-17-CE38-0006).

Le mur d'escalade a été équipé du système Luxov Touch 1 qui utilise une technologie de capteurs pour mesurer le temps de contact et de libération des prises. Tous les essais ont été filmés à 29,97 images par seconde sur $1920\times1080$ pixels par une caméra GoPro 5 (GoPro Inc. 1, San Mateo, CA, USA) qui captait l'intégralité du mur. Le suivi oculaire a été utilisé pour obtenir les coordonnées projetées de la position de la hanche sur le mur en 2D pour chaque image de la vidéo. Le grimpeur portait des lunettes mobiles de suivi oculaire (Tobii Pro Glasses 2© , TobiiAB1, Suède) lors de chaque essai, capables de suivre les mouvements oculaires à une fréquence de 50Hz avec deux caméras sous chaque oeil. Les données de l'eye-tracker mobile et du système Luxov Touch ont été synchronisées en demandant au

participant de regarder une prise (Fig.1 cercle bleu) pendant que l'expérimentateur tapait sur l'emplacement. Ensuite, le temps de la première image de la vidéo de l'eye-tracker montrant le contact du doigt de l'expérimentateur avec la prise a été utilisé comme référence pour synchroniser les deux. Cette synchronisation a été utilisée pour obtenir le temps de décalage du regard, c'est-à-dire la différence de temps entre la dernière visite du regard dans la zone d'intérêt (AOI, 30 cm autour de la prise) de la prise précédente et le contact avec la prise suivante.) Trois directives principales ont été données aux participants avant chaque

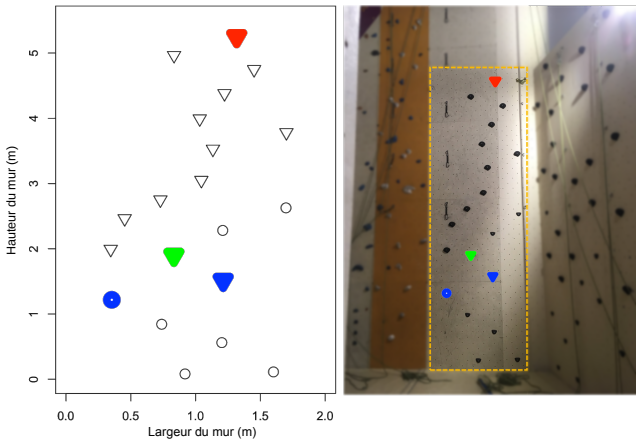



Figure 1: *Conception de l'itinéraire d'entraînement. Les cercles représentent les prises de pied et les triangles les prises de main*

essai, à savoir (i) grimper de la manière la plus fluide possible, c'est-à-dire en minimisant les pauses et les mouvements saccadés du corps ; (ii) utiliser toutes les prises dans un ordre spécifique, de bas en haut ; et (iii) utiliser toutes les prises avec un seul membre à la fois.

La procédure suivante en 9 étapes a été suivie pour chaque essai : (i) la route à gravir est dévoilée (ii) l'eye-tracker mobile est calibré et l'enregistrement commence, (iii) le participant se place à $3m$ devant le mur pour $30s$ de prévisualisation de son trajet. Le participant peut interrompre la prévisualisation quand il le souhaite. Pendant la prévisualisation, les expérimentateurs démarrent l'enregistrement vidéo ; (iv) les participants sont encordés au sommet du mur pour assurer la sécurité pendant les ascensions ; (v) les consignes sont données par l'expérimentateur ; (vi) l'expérimentateur effectue la procédure de synchronisation ; (vii) le participant se place dans la position de départ, en tenant la première prise avec les deux mains (Fig.1 triangle bleu) et ses pieds reposent sur les deux premières prises de pied (bas de la Fig.1) ; (viii) lorsque le participant est prêt et sécurisé, l'expérimentateur annonce qu'il peut commencer à grimper (à partir du triangle vert, Fig.1). L'escalade se termine lorsque le participant saisit la dernière prise (Fig.1 triangle rouge) et reste immobile pendant quelques secondes ; (ix) enfin, le participant redescend et tous les enregistrements sont interrompus.

### 3.2 Analyse statistique

L'analyse statistique suivante est basée sur trois variables, à savoir le décalage du regard (GO), la différence de temps entre deux contacts (TD2H) et la durée de la dernière visite du regard à l'intérieur de la zone d'intérêt (DLGV).

Avant de procéder à l'analyse, il convient de préciser qu'un examen approfondi et un nettoyage des données brutes ont été nécessaires, ce qui n'a permis de garder que 729 essais sur un total de 854. Notre objectif étant d'étudier la dynamique d'apprentissage du grimpeur sur le long terme par le biais du mécanisme DMM, une méthode de clustering non supervisée a été appliquée. Le clustering s'est basé sur les trois variables (renormalisées entre 0 et 1) GO, TD2H et DLGV. Quatre méthodes de clusering ont été testées (*Arbitrary*, *Hierarchical*, *k-Means* et *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)*) et pour chacune de ces méthodes, 3 ou 4 clusters ont été considérés (un dendogramme et la méthode de la silhouette ont été utilisés). La méthode *k-Means* a été choisie, basée sur la distance euclidienne, pour déterminer l'espace d'état $A$. La méthode de clustering ainsi que le nombre de groupes associés ont été sélectionnés sur la base de l'indice de Fisher (FI) défini comme suit :

$$\text{FI} = \frac{\text{Intercluster effect}}{\text{Intra cluster effect}} = \frac{SSB}{SST_w}.$$

Enfin, le DMM sélectionné a été ajusté sur la séquence concaténée des essais de chaque individu à l'aide du package *drimmR* de *R* (Barbu, Vergne, Lothodé, & Seiller, 2021; Barbu, Mavrogiannis, & Vergne, 2023).

## 4 Résultats

Étant donné que FI a obtenu de meilleurs résultats (3,147) pour k = 4, on considère les groupes "a", "b", "c" et "d", illustrés en Figure 2. Un scatterplot 3D interactif des trois variables considérées pour l'implémentation de l'algorithme *k-means* est disponible à https://rpubs.com/ekalligeres/1172114. Les Tableaux 1 et 2 présentent les valeurs du centroïde de chaque groupe ainsi que les écarts-types correspondants. On observe dans le Tableau 1 que (i) le groupe "a" a un TD2H assez long et le deuxième DLGV le plus long ; (ii) le groupe "b" a le GO le plus négatif et le TD2H le plus long ; (iii) le groupe "c" a les valeurs les plus petites de DLGV et TD2H, et ; (iv) le groupe "d" a le DLGV le plus long et un TD2H assez long. On peut en supposer que trois comportements liés à l'efficacité peuvent être distingués. Un comportement moins efficace qui contient les groupes "d" et "b", un comportement moyennement efficace qui contient le groupe "a" et enfin un comportement très efficace qui est représenté par le groupe "c".

Maintenant que nous avons déterminé l'espace d'état A = "a", "b", "c", "d", le DMM optimal choisi (en termes de complexité) est un DMM d'ordre 1 et de degré 3, qui a été appliqué à l'ensemble de la séquence d'essais des 11 individus. Les résultats sont réprésentés dans les Figures 3, 4 et 5. Au début de l'apprentissage, tous les participants utilisent le groupe moins efficace "b" (probabilité la plus élevée) et, du milieu à la fin du processus d'apprentissage, la plupart d'entre eux (c'est-à-dire 3, 4, 5, 6, 9, 10, 11) commencent à

| Cluster | GO | TD2H | DLGV |
|---|---|---|---|
| a (medium) | −0.047 | 0.936 | 2.928 |
| b (less) | −0.241 | 2.100 | 0.960 |
| c (most) | −0.059 | 0.798 | 0.786 |
| d (less) | −0.002 | 1.060 | 6.506 |

Table 1: Valeurs des centroïdes

| Cluster | GO | TD2H | DLGV |
|---|---|---|---|
| a | 0.231 | 0.518 | 0.750 |
| b | 0.344 | 0.751 | 0.570 |
| c | 0.242 | 0.326 | 0.419 |
| d | 0.239 | 0.591 | 1.694 |

Table 2: Valeurs des écarts-types

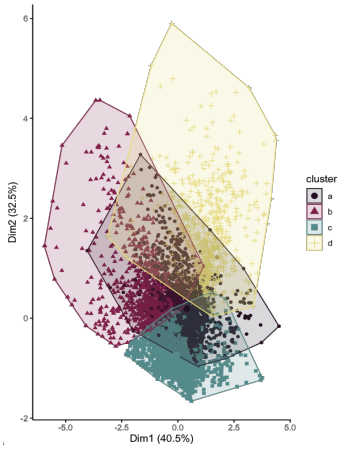



Figure 2: *Clusters "a", "b", "c" et "d". Premier plan de l'ACP*

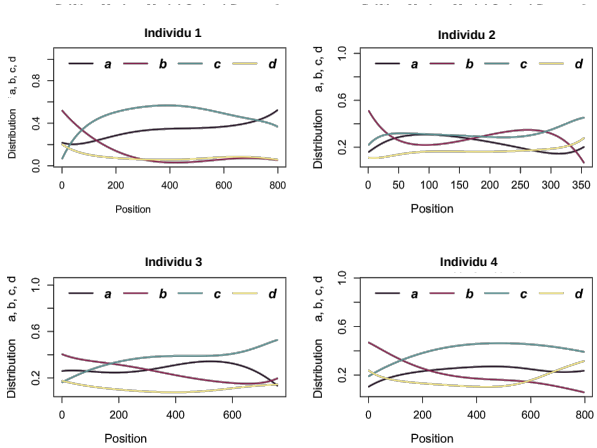



Figure 3: *Dynamique de l'apprentissqage par DMM d'ordre 1 et de degré 3 : individus 1-4*

apprendre et à utiliser le groupe le plus efficace "c". Cependant, certains individus ont un comportement tout à fait différent. En particulier, l'individu 2 adopte assez tôt le groupe "c" le plus efficace, puis l'abandonne pendant plusieurs essais en faveur du groupe "b" et l'adopte de novueau à la fin. Les individus 1 et 7, bien qu'ils adoptent dès le milieu du processus le groupe "c" le plus efficace, semblent utiliser le groupe "a" moyennement efficace avec une probabilité plus élevée à la fin de l'apprentissage. Il en va de même pour l'individu 8 qui, à la fin du processus d'apprentissage, utilise le groupe "d", moins efficace, de façon marginale par rapport au groupe "c", plus efficace.

# Discussion

Les valeurs des centroïdes présentées dans le Tableau 1 ont fourni des informations préliminaires très utiles qui ont ensuite permis de mieux comprendre et interpréter les résultats de la modélisation par DMM. Plus précisément, le groupe "b" présente le GO le plus négatif et le TD2H le plus long, ce qui implique que dans cet état, les individus se concentrent davantage sur l'acquisition d'informations, c'est-à-dire sur l'exploration de leur environnement, à la recherche de nouvelles solutions motrices, plutôt que sur l'utilisation de connaissances a priori afin de continuer à grimper. Le groupe "d" présente la DLGV la plus longue, ce qui reflète le temps nécessaire pour déterminer quelle main doit être utilisée pour saisir la prise (temps de prise de décision et d'enchaînement des actions dans la séquence appropriée) et une TD2H assez longue. Le groupe "a" présente un TD2H assez long et le deuxième

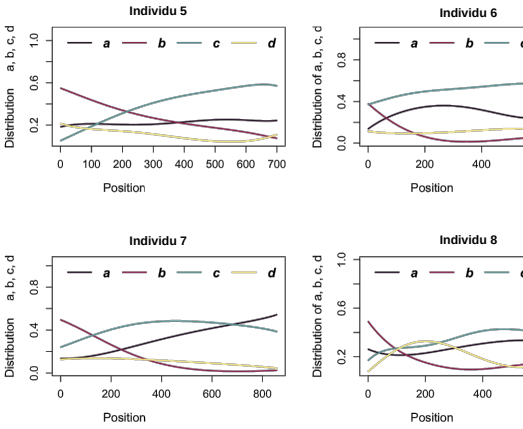



Figure 4: *Dynamique de l'apprentissqage par DMM d'ordre 1 et de degré 3 : individus 5-8*




Figure 5: *Dynamique de l'apprentissqage par DMM d'ordre 1 et de degré 3 : individus 9-11*

DLGV le plus long, ce qui signifie qu'il s'agit d'un état intermédiaire où le comportement d'exploration commence à diminuer pour laisser place à l'exploitation. Cette dernière est définie comme l'utilisation d'un comportement existant dans le répertoire moteur initial ou au début de la pratique. Pendant l'exploitation, les apprenants ne changent pas pour une autre forme de mouvement ou de coordination, mais affinent leur mouvement existant pour le rendre plus efficace. Le rapport entre l'exploration et l'exploitation pourrait expliquer comment les apprenants changent pour une nouvelle pratique ou conservent et affinent les pratiques existantes (Komar, Potdevin, Chollet, & Seifert, 2019). Enfin, dans le groupe "c", l'exploitation a dominé l'exploration puisque les valeurs de DLGV et TD2H ont été considérablement réduites par rapport aux autres groupes.

Lors des premiers essais (Figures 3,4 et 5), tous les participants ont montré une tendance à explorer leur environnement, en utilisant principalement le groupe "b", moins efficace. Ce comportement suggère un contrôle proactif du mouvement pour discerner l'ordre correct d'utilisation de la main droite et de la main gauche. Une telle exploration est considérée comme normale car les individus se concentrent initialement sur la collecte d'informations fonctionnelles pour atteindre la prise suivante. Autour de l'essai 200, certains individus ont évolué vers l'exploitation, favorisant le groupe "c", une tendance qui s'est maintenue jusqu'à la conclusion de l'étude. Des fluctuations ont été observées chez certains individus : par exemple, les individus 1 et 7 ont d'abord favorisé le groupe "b", mais ont adopté un comportement d'exploitation (groupe "c") pendant la majorité de leurs essais. Vers la fin, ils ont fait preuve à la fois d'exploration et d'exploitation. Cette alternance entre l'exploration (groupe "a") et l'exploitation (groupe "c") indique la coexistence de deux comportements visuo-moteurs primaires, caractérisant la bi-stabilité dans l'acquisition des compétences. L'individu 2 a alterné entre l'exploration et l'exploitation, démontrant des tendances à la coexistence de plusieurs comportements sans stabilisation claire. On a pu identifier le concept de "métastabilité": les apprenants opérent de manière intermittente entre l'affinement des comportements visuo-moteurs existants et l'adoption temporaire de nouveaux comportements dus aux exigences de la tâche. L'individu 8 a retardé son départ du comportement d'exploration, adoptant

finalement le groupe "c" après l'avoir temporairement abandonné. Certains individus (4, 5, 6, 9 et 11) ont fait preuve d'un calibrage précis de leur comportement visuo-moteur, en commençant par le groupe "b" moins efficace, en passant au groupe "a" moyennement efficace puis au groupe "c" le plus efficace.

Les conclusions obtenues, facilitées par l'utilisation des DMM, ont mis en évidence, que même si les individus utilisent principalement un comportement visuo-moteur, ils visitent plusieurs solutions stabilisées (reflétant la multistabilité), ou présentent seulement des tendances qui apparaissent ou disparaissent (reflétant la métastabilité). Ces observations s'alignent sur les résultats de Komar et al. (2023), qui indiquent que les individus peuvent présenter des fluctuations dans les probabilités d'exploration et d'exploitation tout au long du processus d'apprentissage.

# Conclusion

Dans ce travail, notre objectif est d'étudier la dynamique d'apprentissage à long-terme d'un grimpeur à travers les DMM. À cette fin, une étude de cas réel a été menée sur 729 essais de 11 individus. En appliquant la méthode de clustering *k-Means* pour obtenir l'espace d'état $A = \{\text{“a”},\text{“b”},\text{“c”},\text{“d”}\}$, trois comportements/modèles liés à l'efficacité ont été distingués : peu, moyennement et très efficace. En outre, en ajustant un DMM d'ordre 1 et de degré 3 sur l'ensemble de la séquence d'essais de chaque individu, nous avons observé que tous les individus "luttaient" au début de l'apprentissage puisqu'ils utilisaient avec une forte probabilité le groupe "b", moins efficace. Tout au long du processus d'apprentissage, le comportement des individus a évolué positivement, conduisant à l'adoption (par la majorité) du groupe "c", le plus efficace. Trois individus ont cependant adopté un comportement différent à la fin du processus. Les résultats ci-dessus soulignent l'utilité des DMM pour comprendre et examiner la dynamique d'apprentissage des grimpeurs et, potentiellement, d'autres disciplines sportives. L'approche par DMM permet de saisir l'évolution à long terme des comportements et de trouver des pratiques qui peuvent être utilisées ultérieurement pour créer des méthodes d'entraînement efficaces et améliorer les performances. D'une manière générale, l'utilisation des DMM en escalade et en sciences du sport ouvre des perspectives prometteuses pour les études futures et les applications dans le monde réel. Les entraîneurs, les athlètes et les chercheurs peuvent développer des interventions et des stratégies ciblées pour améliorer les performances et les résultats de l'entraînement grâce à cette compréhension approfondie des processus d'apprentissage.

# Remerciements

Cette présentation est basée sur Kalligeris (2024).
Ce travail a été financé par la Region Normandie, le Ministère de la Recherche et l'Union Européenne sous un fond CPER-FEDER.

# Bibliographie


Adams, J. (1961). The second facet of forgetting: A review of warm-up decrement. *Psychological Bulletin*, 58, 257–273.

Barbu, V. S., Mavrogiannis, I., & Vergne, N. (2023). dsmmR: Estimation and simulation of drifting semi-markov models. `https://cran.r-project.org/web/packages/dsmmR/index.html`

Barbu, V. S., & Vergne, N. (2019). Reliability and survival analysis for drifting Markov models: Modeling and estimation. *Methodology and Computing in Applied Probability*, 21, 1407–1429.

Barbu, V. S., Vergne, N., Lothodé, C., & Seiller, A. (2021). drimmR: An R package for estimation, simulation, and reliability of drifting Markov models. `https://cran.r-project.org/web/packages/drimmR/index.html`

Draper, N., Giles, D., Schöffl, V., Fuss, F. K., Watts, P., & et al. (2016). Comparative grading scales, statistical analyses, climber descriptors and ability grouping: International rock climbing research association position statement. *Sports Technology*, 8(3-4), 88–94.

Hacques, G., Komar, J., & Seifert, L. (2021). Learning and transfer of perceptual-motor skill: Relationship with gaze and behavioral exploration. *Attention, Perception, & Psychophysics*, 83, 2303–2319.

Hallett, M., & Grafman, J. (1997). Executive function and motor skill learning. *International Review of Neurobiology*, 41, 297–323.

Irion, A. (1948). The relation of "set" to attention. *Psychological Review*, 55, 336–341.

Kalligeris E. N., Barbu V. S. , Hacques G., Seifert L., Vergne N. (2024). Unveiling the Persistent Dynamics of Visual-Motor Skill via Drifting Markov Modeling. *to appear in Nonlinear Dynamics, Psychology, and Life Sciences*.

Komar, J., Potdevin, F., Chollet, D., & Seifert, L. (2019). Between exploitation and exploration of motor behaviours: unpacking the constraints-led approach to foster nonlinear learning in physical education. *Physical Education and Sport Pedagogy*, 24(2), 133–145.

Komar, J., Seifert, L., Vergne, N., & Newell, K. (2023). Narrowing the coordination solution space during motor learning standardizes individual patterns of search strategy but diversifies learning rates. *Scientific Reports*, 13(1).

Newell, K., Kugler, P., Van Emmerik, R., & McDonald, P. (1989). Search strategies and the acquisition of coordination. In S. A. Wallace (Ed.), *Perspectives on the coordination of movement* (Vol. 61, pp. 85–122). Amsterdam: North-Holland.

Vergne, N. (2008). Drifting Markov models with polynomial drift and applications to DNA sequences. *Statistical Applications in Genetics and Molecular Biology*, 7(1). doi: doi:10.2202/1544-6115.1326

# Session groupe Environnement et Statistique

# L'IA POUR LES PRÉVISIONS METEOROLOGIQUES ET CLIMATIQUES : ETAT DES LIEUX ET PERSPECTIVES

Laure Raynaud
CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
laure.raynaud@meteo.fr
*Avec des contributions scientifiques de plusieurs collègues.*

**Résumé**

**xxxxx**

L'intelligence artificielle (IA), et en particulier les méthodes d'apprentissage profond utilisant les réseaux de neurones, a été exploitée avec succès dans un grand nombre d'applications ces dernières années. La prévision météorologique et climatique ne fait pas exception, les utilisations de l'IA dans ce domaine sont potentiellement nombreuses et pourraient conduire à des avancées méthodologiques majeures, associées à des gains significatifs en performance et en qualité. Toutes les étapes de la chaîne de production des prévisions sont concernées, de l'assimilation des données au post-traitement des prévisions, en passant par l'élaboration des modèles de prévision, de la très courte échéance aux projections climatiques.

Suite à cette nouvelle tendance, Météo-France a commencé à étudier différentes pistes pour intégrer l'IA dans le calcul et l'exploitation des prévisions météorologiques et climatiques. Qu'ils soient exploratoires ou plus proches d'une utilisation opérationnelle, ces travaux illustrent le potentiel de l'IA pour certaines de nos activités, et ouvrent de nouvelles questions et défis quant au statut de l'IA dans les prochaines évolutions des systèmes de prévision.

Nous décrivons brièvement quelques recherches en cours à Météo-France pour exploiter l'IA en tant que nouvel outil pour la modélisation atmosphérique. Différentes stratégies sont discutées, de l'hybridation entre les modèles physiques de prévision et l'IA, jusqu'au développement d'émulateurs 'tout IA' annonçant un potentiel changement de paradigme pour la prévision météo-climatique.

**Mots clés**

Prévision météorologique, apprentissage statistique, réseaux de neurones

**Abstract**

**xxxxx**

Artificial Intelligence (AI), and in particular deep learning methods based on neural networks, has been successfully leveraged in a growing number of areas in recent years. The applications of AI to weather and climate forecasting are potentially numerous, and could lead to major methodological breakthroughs, associated with significant gains in performance and quality. All stages in the forecast processing chain may be concerned, from data assimilation to forecasts analysis, as well as the development of models, from nowcasting to climate projections.

Following this new trend, Météo-France started investigating the use of AI for different aspects of weather and climate predictions. Whether exploratory or closer to operationalization, these works illustrate the potential of AI for some of our activities, and open up new questions and challenges regarding the status of AI in the next evolutions of forecasting systems.

We briefly describe and discuss ongoing research at Météo-France to leverage AI as a new component for atmospheric modelling, from soft hybridization of AI and physics-based weather forecasting models to the more disruptive full data-driven weather forecasting emulators.

**Keywords**

Weather prediction, machine leaning, neural networks

# 1 Introduction

La prévision météorologique est le résultat d'une séquence d'étapes complexes dont l'élément central est le modèle de prévision. Le développement de ces modèles, initié dès les années 1950, s'appuie sur la connaissance experte des lois physiques qui gouvernent l'évolution de l'atmosphère. Plusieurs fois par jour, la résolution numérique des équations du modèle, sur des grilles plus ou moins fines, permet de produire des prévisions pour les prochaines heures et jours.

La prévision numérique du temps est un domaine en évolution constante. L'intégration de nouvelles observations et de représentations plus précises des processus physiques, en lien avec l'augmentation des ressources de calcul, permet d'améliorer régulièrement la qualité des prévisions et l'anticipation des événements à fort impact. Néanmoins, la modélisation reste une approximation du système réel, limitée par notre compréhension des processus en jeu et par les contraintes computationnelles.

L'utilisation de l'intelligence artificielle (IA) en météorologie n'est pas nouvelle. Dès les années 1990, les techniques d'IA ont permis des développements novateurs dans le post-traitement statistique des prévisions météorologiques. Diverses applications été développées, notamment à Météo-France, pour réduire les erreurs systématiques des prévisions. Par exemple, le traitement par forêts aléatoires des prévisions de pluies extrêmes permet de les corriger pour se rapprocher des valeurs observées (Taillardat et al., 2019).

En revanche, ce n'est que récemment que l'utilisation de l'IA s'est étendue au coeur de la modélisation atmosphérique. C'est en particulier la capacité de l'IA à apprendre des relations complexes et à opérer très rapidement dans sa phase d'inférence qui a motivé l'utilisation de ces approches pour améliorer les modèles de prévision et réduire leur coût de calcul. La complémentarité des approches physique et IA a initialement motivé le développement de systèmes de prévision « hybrides », combinant la modélisation physique classique et l'IA. Il s'agit par exemple de remplacer les éléments les plus coûteux ou les moins bien représentés d'un modèle physique par un algorithme d'IA. D'autres travaux se sont penchés sur la possibilité d'exploiter l'IA pour améliorer certaines caractéristiques des prévisions (finesse de l'échantillonnage spatial ou probabiliste par exemple), et in fine leur qualité, à moindre coût. Une nouvelle étape a été franchie il y a peu par plusieurs équipes de recherche, qui ont proposé de remplacer complètement le modèle physique par un modèle d'IA. Le travail précurseur de Dueben et Bauer (2018) avait jeté les bases d'une prévision purement statistique, mais en émettant de sérieuses réserves sur sa capacité à rivaliser un jour avec la prévision physique. Cinq ans plus tard, les progrès ont été bien plus rapides qu'attendus, et une partie de ces doutes est levée.

L'étude de Ravuri et al. (2021) est une des premières à montrer qu'un modèle 'tout IA' (de type génératif) peut réaliser des prévisions de précipitations réalistes à très courte échéance. Depuis 2022, une succession de travaux attaque le problème de la prévision globale à moyenne échéance (Bi et al., 2022 ; Lam et al., 2023, Lguensat et al., 2023). Ces modèles d'IA, dans la suite appelés *émulateurs,* sont représentés par exemple par les systèmes Pangu-Weather ou GraphCast. Entraînés sur plus de 40 ans de données, ils sont désormais compétitifs sur plusieurs aspects avec le modèle physique du Centre Européen de Prévision Météorologique à Moyen Terme (CEPMMT), considéré comme le meilleur modèle de prévision opérationnel actuellement (Ben Bouallègue et al., 2023). Cette percée très rapide des émulateurs est d'autant plus surprenante que, pour la première fois, elle n'est pas portée par les services météorologiques nationaux mais par des grandes entreprises internationales telles que Nvidia, Huawei, DeepMind ou Microsoft. C'est donc un potentiel double changement de paradigme pour la communauté météorologique.

La suite du document présente quelques travaux engagés à Météo-France pour exploiter l'IA, et en particulier les méthodes neuronales, dans le processus de prévision. Une analyse succincte des émulateurs existants et des pistes de recherche qu'ils ouvrent est également proposée.

# 2 L'IA pour la prévision à Météo-France : quelques premières réalisations

## 2.1 L'IA pour améliorer la physique des modèles

Les paramétrisations physiques, qui simulent les effets des processus sous-maille tels que le rayonnement, la convection ou la turbulence, sont actuellement parmi les composants les plus coûteux d'un modèle, ainsi qu'une des principales sources d'incertitude des prévisions météorologiques et climatiques. Plusieurs travaux ont commencé à examiner la possibilité de remplacer tout ou partie de ces paramétrisations par des algorithmes d'IA. A titre d'exemple citons le travail de Balogh (2022), qui a réalisé une simulation d'une durée d'un an avec le modèle ARPEGE-Climat où le schéma de convection profonde a été remplacé par des réseaux de neurones de type récurrents (Figure 1).

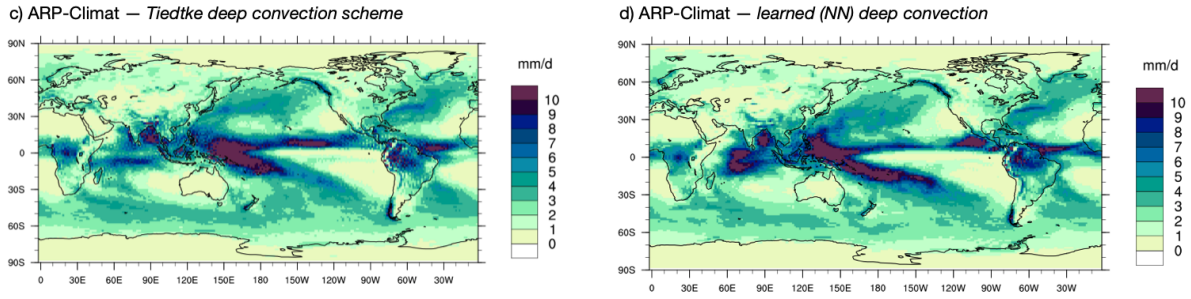



*Figure 1: Précipitations annuelles moyennes prévues par la modèle ARPEGE-Climat. A gauche, prévision du modèle physique. A droite, prévision d'un modèle « hybride », où la paramétrisation de la convection profonde est remplacée par un réseau de neurones. Source : B. Balogh.*

## 2.2 L'IA pour améliorer l'échantillonnage spatial : la super-résolution appliquée aux prévisions météo

Augmenter la résolution du modèle de prévision permet de mieux décrire les phénomènes météorologiques de fine échelle. C'est particulièrement important pour la prévision d'événements comme les orages, le brouillard, l'îlot de chaleur urbain ; mais au prix d'une augmentation significative, voire rédhibitoire, du coût de calcul. La descente d'échelle statistique, que l'on peut grossièrement comparer à la super-résolution, est une alternative à l'augmentation de résolution du modèle, qui consiste à apprendre une relation statistique entre les prévisions à basse résolution et les prévisions à plus haute résolution. Il est ainsi possible de simuler des prévisions haute résolution en appliquant a posteriori cette relation aux prévisions d'un modèle à basse résolution. De nombreuses études se sont penchées sur la capacité des réseaux de neurones, notamment convolutionnels, à résoudre ce problème, avec des résultats très prometteurs. On peut mentionner les travaux de Doury et al. (2022), qui proposent l'application d'une architecture de type U-Net pour la descente d'échelle de simulations climatiques (Figure 2).

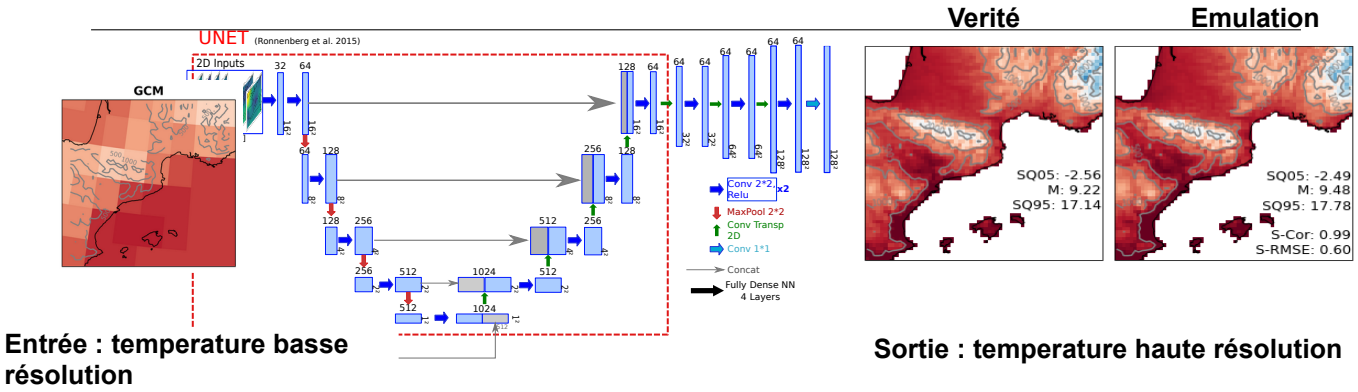



*Figure 2 : Principe de la descente d'échelle statistique pour la prévision météorologique ou climatique. Source : A. Doury.*

## 2.3 L'IA pour améliorer l'échantillonnage probabiliste : les méthodes génératives appliquées à la météo

La prévision d'ensemble est utilisée pour caractériser les différents scénarios d'évolution possibles grâce à la réalisation en parallèle de plusieurs prévisions. Devenue un élement incontournable de nombreux services de prévision, la prévision d'ensemble n'en reste pas moins fortement contrainte par les ressources de calcul disponibles, en particulier en ce qui concerne le nombre de réalisations (aussi appelées « membres »). Les prévisions d'ensemble actuellement opérationnelles n'utilisent pas plus de 50 membres, alors qu'une estimation précise des distributions de probabilité en requiert plusieurs centaines ou milliers. L'IA pourrait-elle permettre de générer des membres supplémentaires en se substituant au modèle de prévision ? Les travaux de Brochet et al. (2023) apportent de premiers éléments de réponse encourageants. S'appuyant sur une architecture de type Generative Adversarial Network (GAN), ils montrent qu'il est possible de produire des champs météorologiques réalistes, et ouvrent ainsi la voie à des prévisions d'ensemble hybrides de plusieurs dizaines voire centaines de membres (Figure 3).

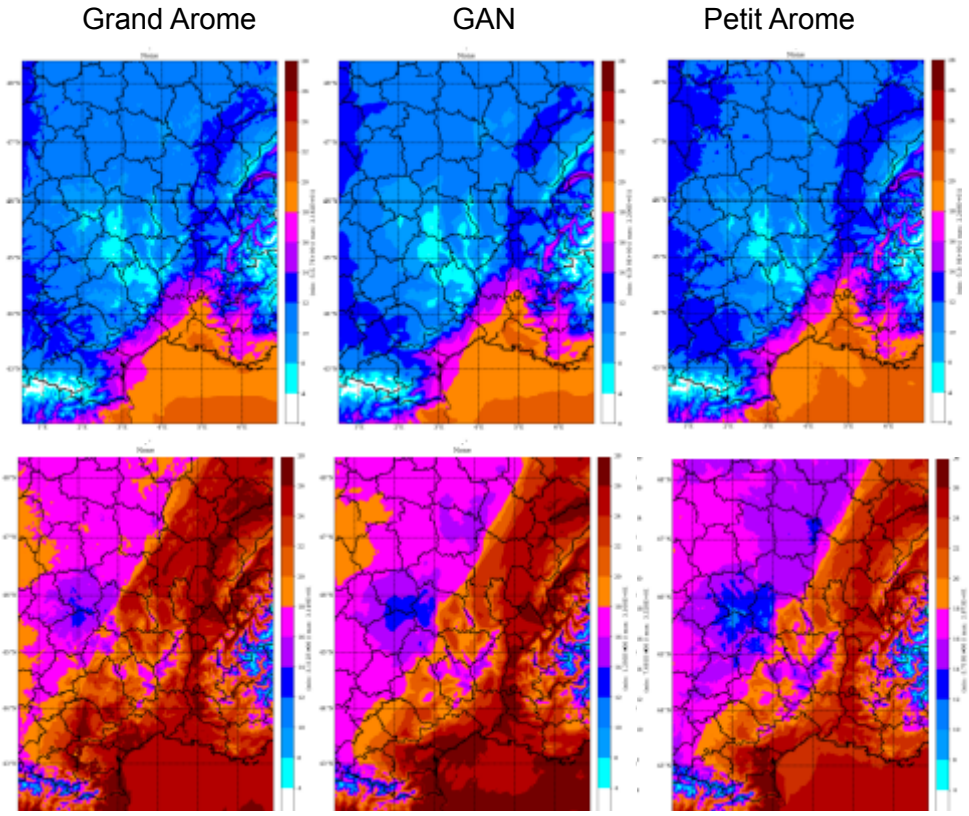


*Figure 3 : Comparaison des quantiles min (ligne du haut) et max (ligne du bas) des distributions de température échantillonnées par trois ensembles. A gauche (resp. droite) un ensemble de 875 (resp. 16) membres utilisant le modèle de prévision Arome, au milieu un ensemble de 875 membres générés par un GAN (conditionné aux 16 membres Arome). Le GAN permet d'étendre la distribution du petit ensemble Arome et de s'approcher de celle du grand ensemble Arome. Les membres générés par le GAN sont par ailleurs physiquement cohérents (Brochet et al., 2023). Source : C. Brochet, G. Moldovan.*

## 2.4 Les émulateurs sont-ils l'avenir de la prévision météo ?

Les émulateurs publiés récemment vont un cran plus loin que les réalisations présentées précédemment, en proposant une alternative complète aux modèles de prévision physiques, et en exploitant des architectures neuronales complexes, basées principalement sur les vision transformers, ou une représentation sous forme de graphes. Dans toutes ces expériences, la base d'apprentissage est la même, et correspond à la reconstruction de l'état de l'atmosphère depuis 1940 jusqu'à aujourd'hui, heure par heure, à une résolution horizontale de 30 km sur tout le globe. Ce jeu de données est une formidable source d'information sur notre système Terre et permet notamment de suivre l'évolution du climat, même si sa résolution spatiale est loin des standards des modèles de prévision en opération, qui atteignent aujourd'hui des résolutions globales inférieures à 10km et régionalement des résolutions kilométriques, voire hectométriques. Ces démonstrateurs offrent une représentation encore très partielle de l'atmosphère, avec un nombre limité de variables prévues, une résolution spatiale loin de l'état de l'art, et des cohérences physiques imparfaites, mais ils ouvrent un nouveau pan de recherche avec de multiples questions scientifiques et techniques, et des opportunités nouvelles pour la prévision opérationelle.
La Figure 4 présente les performances de ces émulateurs, avec des scores objectifs meilleurs que ceux des prévisions du CEPMMT pour certaines variables et échéances. Ces émulateurs ont également montré leur capacité à simuler des événements à fort impact, tel qu'illustré ici sur la tempête Ciaran. Ces résultats, bien qu'encourageants, doivent encore être consolidés avant de pouvoir considérer ces émulateurs suffisamment robustes et fiables pour un usage en conditions opérationnelles.

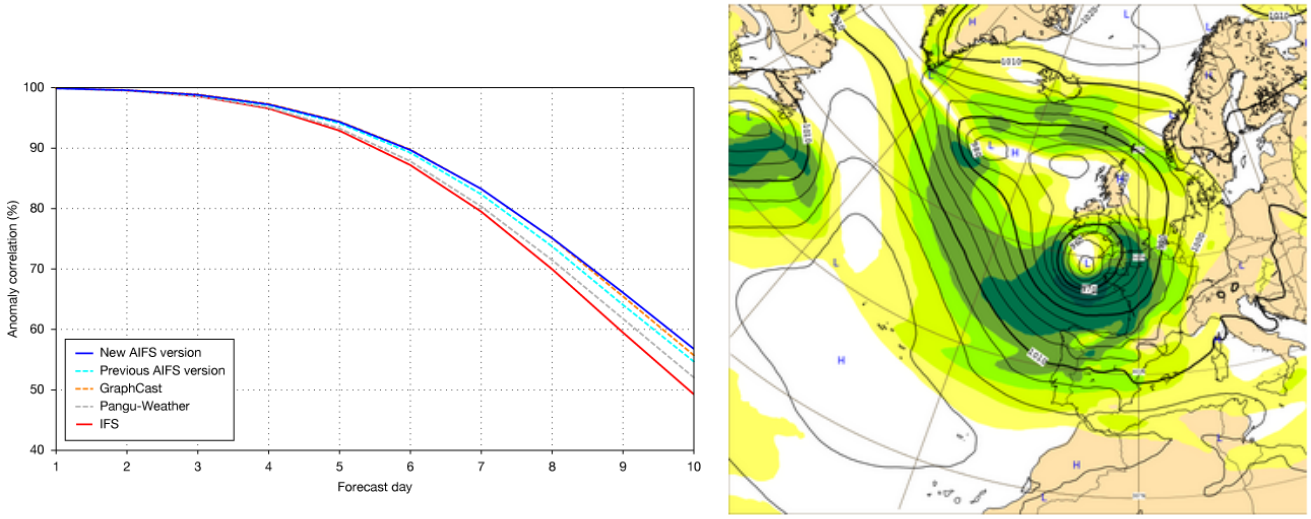



*Figure 4 : Gauche : score de corrélation d'anomalie pour le géopotentiel à 500hPa en fonction de l'échéance de prévision (source : CEPMMT). Plusieurs émulateurs IA sont comparés à la prévision du CEPMMT (en rouge). Plus le score est proche de 100% meilleure est la prévision. Droite : prévision de l'émulateur Pangu Weather pour le cas de la tempête Ciaran (validité 2 novembre 2023 à 0 UTC), pression au niveau de la mer (isolignes) et force du vent à 850hPa (plages de couleurs), source : M. Pardé.*

Malgré des faiblesses bien identifiées, la démonstration est faite qu'il est possible de prévoir une partie des paramètres météorologiques, avec une certaine qualité, et dans des temps très courts. Les temps de production sont de quelques secondes ou minutes, soit bien inférieurs à ceux des modèles de prévision physiques qui se comptent en dizaines de minutes, voire plus. Il demeure qu'en amont, il faut disposer de la capacité à réaliser la phase d'apprentissage qui, elle, peut prendre des temps longs, qui se comptent en semaines ou mois, mais qui sont hors du chemin critique de la production des prévisions en temps réel.

Pour que ces émulateurs deviennent de nouveaux outils exploitables pour la prévision du temps opérationnelle, et plus généralement pour toutes les applications nécessitant des données météorologiques, il reste de nombreux verrous à lever. Le premier enjeu est de développer des émulateurs adaptés aux besoins des usagers, entraînés sur des données à très haute résolution spatiale, et capables de prévoir les variables météorologiques d'intérêt et les incertitudes associées. Cela soulève la question de la disponibilité et de l'accessibilité de ces jeux de données, et de la capacité à mobiliser des ressources de calcul conséquentes pour des entraînements pouvant atteindre plusieurs semaines. Les émulateurs actuels ont pour la plupart été conçus pour des applications spécifiques. Le développement d'émulateurs plus génériques, de type modèles de fondation, est certainement la suite à envisager pour couvrir efficacement une large gamme d'objectifs (Nguyen et al., 2023). Enfin, une étape importante à franchir sera l'apprentissage des émulateurs sur des données multi-sources (modèles, observations), avec des qualités et des couvertures spatio-temporelles hétérogènes.
Le second enjeu est la mise au point d'outils et de diagnostics d'interprétabilibité et d'explicabilité de ces émulateurs. A l'instar des modèles physiques, il est légitime de pouvoir déterminer si les émulateurs ont produit une bonne prévision pour les bonnes raisons ou, en cas de mauvaises prévisions, quels composants de l'architecture sont en cause. Une perspective sous-jacente est le développement de réseaux de neurones informés par la physique, pour forcer les émulateurs à produire des solutions physiquement cohérentes.

# 3 Conclusions

Si le développement incrémental de la prévision numérique du temps est souvent qualifié de 'révolution lente', c'est une révolution beaucoup plus rapide qui semble se mettre en marche avec l'avènement des premiers modèles de prévision par IA. C'est aussi un nouveau pan de recherche qui s'ouvre devant les services météorologiques, avec de nouveaux enjeux scientifiques et techniques. Cela ne doit néanmoins pas faire perdre de vue la poursuite des travaux d'amélioration des modèles de prévision physiques. Il ne s'agit pas, à ce stade, de remplacer l'un par l'autre, mais d'exploiter la complémentarité de ces deux approches.

**Bibliographie**

Balogh B.**,** 2022**.** Vers une utilisation de l'Intelligence Artificielle dans un modèle numérique de climat. Thèse de doctorat en Océan, atmosphère, climat. Toulouse INPT.

Bi K., L. Xie, H. Zhang, X. Chen, X. Gu, Q. Tian, 2022 : Pangu-Weather - A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast, arXiv preprint arXiv:2211.02556.

Ben Bouallègue Z. et al., 2023 : The rise of data-driven weather forecasting, a first statistical assessment of machine learning-based weather forecasts in an operational-like context

Brochet, C., L. Raynaud, N. Thome, M. Plu, and C. Rambour, 2023: Multivariate Emulation of Kilometer-Scale Numerical Weather Predictions with Generative Adversarial Networks: A Proof of Concept. Artif. Intell. Earth Syst., 2, 230006, https://doi.org/10.1175/AIES-D-23-0006.1

Doury, A., Somot, S., Gadat, S. et al. Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach. Clim Dyn (2022). https://doi.org/10.1007/s00382-022-06343-98

Dueben, P. D. et Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geosci. Model Dev., 11, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018, 2018.

Lam R. *et al.*, Learning skillful medium-range global weather forecasting. *Science*, DOI:10.1126/science.adi2336

Lguensat R., 2023 : Les nouveaux modèles de prévision météorologique basés sur l'intelligence artificielle : opportunité ou menace ? La Météorologie n°121 mai 2023 pp 11-15

Nguyen T., J. Brandstetter, A. Kapoor, J. K. Gupta and A. Grover, 2023 : ClimaX - A foundation model for weather and climate. arXiv preprint arXiv:2301.10343v2, 41p, 24 Jan 2023.

Ravuri, S., Lenc, K., Willson, M. *et al.* Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**, 672–677 (2021). https://doi.org/10.1038/s41586-021-03854-z

Taillardat, M., A. Fougères, P. Naveau, and O. Mestre, 2019: Forest-Based and Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting. *Wea. Forecasting*, **34**, 617–634, https://doi.org/10.1175/WAF-D-18-0149.1.

# INFLUENCE DU CLIMAT SUR L'EXPRESSION DES SYMPTOMES D'UNE MALADIE VASCULAIRE DE LA VIGNE

Chloé Delmas[1*], Lucas Etienne[1], Thibaut Fréjaville[1], Davide Martinetti[2], Frédéric Fabre[1], Elise Frank[3], Pascal Lecomte[1], Lucie Michel[3], Valérie Bonnardot[4] & Lucia Guérin-Dubrana[1]

*[1] INRAE, ISVV, Bordeaux Sciences Agro, Santé et Agroécologie du Vignoble, 33140 Villenave d'Ornon, France*
*[2] INRAE, Biostatistiques et Processus Spatiaux, 84000 Avignon, France*
*[3]Plateforme ESV, INRAE, Biostatistiques et Processus Spatiaux, 84914 Avignon, France*
*[4] CNRS, Université Rennes 2, Littoral Environnement Télédétection Géomatique, 35045 Rennes, France*
**chloe.delmas@inrae.fr*

**Résumé.** Les interactions entre facteurs abiotiques (comme le climat) et biotiques (comme les agents pathogènes) sont à l'origine d'un processus de dépérissement impactant particulièrement les plantes pérennes dans le contexte du changement climatique. La viticulture mondiale est largement impactée par ce processus, notamment du fait de maladies du bois qui pèsent lourdement sur la longévité des vignobles. Parmi ces maladies, l'esca se distingue comme l'une des plus préoccupantes. Cette maladie vasculaire résulte d'une association complexe de divers agents fongiques et se caractérise par l'expression de symptômes foliaires. Ses déterminants sont toutefois largement méconnus. Ici, nous proposons d'étudier le rôle d'un ensemble de facteurs, notamment climatiques, sur l'incidence de l'esca, à de larges échelles spatiales et temporelles, en utilisant les outils de la modélisation statistique.

Pour étudier la dynamique intra-saisonnière de l'esca ainsi que l'incidence annuelle des symptômes en relation avec le climat en France, deux types d'approches de modélisation ont été utilisées, toutes deux reposant sur les variables climatiques de la base de données SAFRAN (Météo-France) estimées sur des mailles horizontales de 8 x 8 km. D'une part, une étude des observations de symptômes foliaires de l'esca sur 50 parcelles plantés avec 11 cépages et suivies entre 2003 et 2021 a été réalisée pour analyser l'incidence hebdomadaire et la phénologie de l'esca au fil des saisons, en utilisant un ensemble de modèles linéaires généralisés mixtes (*GLMM*) avec des approches fréquentistes. D'autre part, une étude sur 481 parcelles plantées avec 7 cépages et suivies sur la même période a permis d'explorer l'effet du climat en relation avec la phénologie de la vigne (calculs d'indicateurs écoclimatiques) sur l'incidence annuelle des symptômes, en utilisant des *GLMM* avec des approches bayésiennes reposant sur la méthode *Integrated nested Laplace approximation* (INLA). Ces approches complémentaires ont mis en évidence le rôle du climat dans l'incidence de l'esca à différents niveaux temporels (intra- et inter-annuelle).

A l'échelle intra-annuelle, nos résultats suggèrent qu'un climat plus chaud, dans les deux mois précédents l'expression, conduit à des symptômes d'esca plus précoces durant la saison, mais diminue la proportion de nouvelles vignes symptomatiques de l'escax x du fait d'une sécheresse printanière accrue et d'une forte corrélation positive entre les températures au printemps et le VPD (déficit de vapeur d'eau). A l'échelle inter-annuelle, nos résultats suggèrent que le climat pendant la saison de croissance et les périodes d'expression des symptômes foliaires ont fortement influencé l'incidence annuelle de l'esca des différents cépages étudiés, en particulier la température moyenne, la disponibilité en eau du sol, l'évapotranspiration et le VPD. Ces résultats offrent des perspectives intéressantes pour mieux appréhender les interactions tripartites entre plantes, climat et maladies. Ils permettront de

mieux anticiper le risque de dépérissement dans les décennies à venir.

**Mots-clés.** Climat - épidémiologie végétale - esca – INLA – vigne – santé des plantes – modèles mixtes – SAFRAN - statistique bayésienne

**Abstract.** The decline of perennial plants due to interactions between abiotic factors, such as climate, and biotic factors, such as pathogens, is a significant issue in the context of climate change. Plant dieback is having a significant impact on wine production worldwide, particularly due to wood diseases that affect the longevity of vineyards. Esca is one of the most concerning diseases. This vascular disease involves a complex community of various fungal pathogens and is characterized by typical foliar symptoms. The factors influencing foliar symptom incidence are still poorly understood. Here, we propose to investigate the impact of various factors, including climate, on esca incidence on a large spatial and temporal scale, using statistical modelling.

To investigate the intra-seasonal dynamics of esca and the annual expression of symptoms in relation to climate in France, two modelling approaches were employed. Both approaches used climate variables from the SAFRAN (Météo France) database estimated on 8x8km horizontal grids. Firstly, a study of esca leaf symptom observations in 50 plots planted with 11 varieties and monitored between 2003 and 2021 was conducted to analyze the weekly incidence and phenology of esca over the seasons, using an ensemble of linear generalized mixed models (*GLMM*) fitted with frequentist approaches. Secondly, a study of 481 plots planted with 7 cultivars and monitored during the same period investigated the relationship between climate and vine phenology (ecoclimatic indicators) on the yearly incidence of symptoms, using GLMM fitted with Bayesian approaches and the method Integrated nested Laplace approximation (INLA). These complementary methods have revealed the role of climate in esca incidence at various temporal levels (intra- and inter-annual).

On an intra-annual scale, our results suggest that a warmer climate in the two months preceding expression leads to esca symptoms appearing earlier in the season, but reduces the proportion of new symptomatic vines due to increased spring drought and a strong positive correlation between spring temperatures and VPD (vapor pressure deficit). On an inter-annual scale, the climate during the growing season and leaf symptom expression periods strongly affected the annual incidence of esca of the different cultivars studied, in particular the average temperature, soil water availability, evapotranspiration, and VPD. These findings provide new insights for comprehending the relationships between plants, climate, and diseases, allowing us to anticipate the risk of vine decline in the future decades.

**Keywords.** Climate - plant epidemiology - esca - INLA - vine - plant health - mixed models - SAFRAN - Bayesian statistics

# 1. Contexte de l’étude et objectifs

## 1.1 Dépérissement du vignoble et changement climatique

Le changement climatique pourrait être favorable aux cycles biologiques de nombreux champignons pathogènes augmentant ainsi la pression qu’ils exercent sur les cultures des climats tempérés (Chaloner et al. 2021). Cependant, l’augmentation de l’intensité et de la fréquence des sécheresses pourrait contrebalancer les effets favorables du réchauffement sur les agents pathogènes et la productivité des cultures (Torres-Ruiz et al. 2024). Par conséquent, la compréhension des relations entre les conditions climatiques, les agents pathogènes et le développement des maladies s’avère cruciale afin d’anticiper les effets du changement climatique sur la santé des cultures. Notamment, de nombreuses plantes pérennes subissent un dépérissement global causé par des facteurs biotiques, abiotiques et leurs interactions (Allen et al. 2010 ; Hammond et al. 2022).

Les maladies du bois de la vigne, en particulier l’esca, constituent actuellement l'un des défis majeurs de la viticulture, étant une des causes du dépérissement des vignobles (Gramaje et al. 2018). L'esca est une maladie complexe qui implique plusieurs champignons pathogènes pouvant entraîner la mort des vignes. L'incidence des symptômes foliaires d’esca (Figure 1) a augmenté depuis le début du 21e siècle, en partie en raison de l'abandon progressif de l’arsénite de sodium qui était jusqu’en 2001, année de son interdiction, le seul traitement chimique disponible (Bertsch et al. 2013).

Figure 1. Symptômes foliaires d’esca sur le cépage de vigne Sauvignon blanc (Villenave d’Ornon, France) ©Delmas

La variabilité des symptômes foliaires rend la maladie difficile à comprendre, avec des plantes présentant parfois des symptômes une année mais pas les suivantes (Dewasme et al. 2022). La phénologie de cette maladie (dates d’apparition des symptômes) et les variations intra- et inter-saisonnière de son expression restent peu étudiées (mais voir Lecomte et al. 2024), malgré l’importance de ces caractéristiques pour la gestion des vignobles. Les variations climatiques (journalières, saisonnières et annuelles), à plus ou moins long terme, de par leurs effets directs sur la biologie des champignons et des plantes infectées, jouent très probablement un rôle clé dans les variations d’incidence des symptômes foliaires d’esca.

L’étude du rôle des facteurs climatiques dans la santé des plantes nécessite la collecte de données d’épidémiosurveillance à l’échelle des territoires et sur de nombreuses années. Les données d’incidence des symptômes, observés à l’échelle des plantes (Nutter et al. 2006), sont souvent complexes à modéliser chez la vigne du fait de la pluralité des facteurs en jeu (année, cépage, âge et localisation des parcelles, taille d’échantillon, méthode de sélection des parcelles suivies, pratiques culturales).

## 1.2 Objectifs

L'enjeu de ce travail est d'explorer le rôle du climat dans l'incidence de l'esca au cours des saisons de végétation et des vingt dernières années à l'échelle nationale. Pour cela deux objectifs ont été réalisés indépendamment :

(1) Étudier le rôle du climat dans l'incidence intra-saisonnière des symptômes foliaires de l'esca et la phénologie de la maladie (date où 50% de l'incidence finale de la saison est atteinte) dans un réseau de vignobles du sud de la France à l'aide d'un ensemble de modèles mixtes fréquentistes ;

(2) Étudier l'influence d'indicateurs écoclimatiques (variables climatiques calculées en fonction de la phénologie des cépages de vigne) sur l'expression des symptômes foliaires de l'esca en fin de saison (incidence annuelle) à l'aide d'une base de données de surveillance nationale et de modèles bayésiens (*Integrated nested Laplace approximation* - INLA).

# 2. Méthodologie : intégrer les données épidémiologiques et indicateurs climatiques

## 2.1 Données épidémiologiques

Pour l'objectif 1, le jeu de données « épidémiologie » consiste en des notations quasi-hebdomadaires de symptômes foliaires d'esca à l'échelle du cep de vigne réalisées dans le Sud-Ouest (voir Lecomte et al. 2024) et dans le Sud-Est de la France. Cinquante vignobles et 11 cépages ont été suivis entre 2003 et 2021. Pour chaque couple parcelle-année, l'incidence de l'esca, correspondant à la proportion de ceps de vigne nouvellement symptomatiques par date de notation et par site, a été estimée par semaine ainsi que la date à laquelle 50% de l'incidence finale cumulée était atteinte.

Pour l'objectif 2, le jeu de données utilisé est issu de la surveillance nationale des vignobles (observatoire maladies du bois). Ces données d'observation ont été regroupées et homogénéisées pour constituer un unique set de données (projet CLIMESCA) hébergé dans le système d'information de la Plateforme d'Épidémiosurveillance en Santé Végétale. Les données présentées ici regroupent les incidences annuelles (proportion de ceps de vigne symptomatiques au 31 août, sur un nombre N de ceps observés, en moyenne 400) de 481 parcelles et 7 cépages suivis entre 2003 et 2021 en Alsace-Lorraine, Bordelais, Bourgogne, Charentes, Champagne et Val de Loire.

## 2.2 Données climatiques

Pour décrire le contexte climatique de chaque parcelle chaque année, faute de couverture spatiale suffisante par le réseau national de stations météorologiques, le jeu de données météorologiques interpolées SAFRAN (Système d'analyse fournissant des renseignements atmosphériques à la neige) a été utilisé. Les données climatiques SAFRAN proviennent de Météo-France et ont été téléchargées via la plateforme SICLIMA développée par AgroClim-INRAE. Ces données climatiques journalières historiques sont issues de réanalyse spatialisées à une résolution horizontale de 8 x 8 km (Vidal et al. 2010). Il faut ici noter que ces données comportent bien sûr certaines limites et notamment un « biais froid » en raison d'une sous-

estimation du rayonnement infra-rouge (Le Moigne et al. 2020).

Plusieurs variables climatiques ont été sélectionnées comme l'humidité relative, l'évapotranspiration potentielle et réelle, les précipitations, les températures ou encore l'indice d'humidité du sol et ont été extraites pour la période d'intérêt 2003-2021 pour chaque parcelle étudiée (objectif 1). Les valeurs moyennes, ou minimales et maximales de ces variables ont été calculées pour différentes périodes précédant les incidences hebdomadaires de l'esca au vignoble (de 7 jours à 4 mois), afin de modéliser le rôle du climat dans l'expression de l'esca.

Pour l'objectif (2), les variables climatiques sont calculées en fonction de la phénologie de chaque cépage de vigne (indicateurs écoclimatiques). Les calculs d'indicateurs ont été réalisés à partir de la plateforme SICLIMA (AgroClim-INRAE) qui intègre la bibliothèque d'indicateurs développée par Garcia de Cortazar-Atauri and Maury (2019) et qui utilise l'approche des indicateurs écoclimatiques développée par Caubel et al., 2015. Le modèle permet le calcul de stades de phénologie chaque année sur une gamme de cépages et les indicateurs écoclimatiques sont calculés entre ces stades phénologiques (https://agroclim.inrae.fr/siclima/help/decouvrir/calcul.html).

### 2.3 Modélisations statistiques

Les relations entre le climat et l'évolution de l'esca au cours de la saison (objectif 1) ont été testées à l'aide d'une approche d'ensemble de modèles linéaires à effets mixtes (Fréjaville et al. 2020). Les variables climatiques ont été définies comme des effets fixes. Le site (vignoble) a été inclus comme covariable à effet aléatoire, de même que l'âge de la plante, le cépage (n=11) et la région d'étude (Sud-Ouest vs. Sud-Est). Avant de tester différentes combinaisons de variables climatiques pour obtenir un ensemble de modèles, nous avons sélectionné les variables climatiques et les différentes périodes d'influence du climat (7, 15, 30, 60 ou 90 jours précédant l'expression des symptômes) en ajustant un modèle linéaire à effets mixtes, avec ou sans terme quadratique et en le comparant à un modèle dit « nul » (sans variable climatique) en utilisant le critère d'information d'Akaike de second ordre (AICc).

Concernant l'effet des indicateurs écoclimatiques sur l'incidence annuelle de l'esca en France (objectif 2), nous avons utilisé un modèle linéaire mixte généralisé (GLMM) dans un cadre bayésien avec les méthodes INLA (integrated nested Laplace approximation ; Rue et al., 2017). La probabilité de présence de l'esca sur la parcelle a été modélisée à l'aide d'une distribution binomiale avec une fonction de lien logit. Les différents cofacteurs (l'année en interaction avec le cépage et l'âge de la parcelle) ont été pris en compte afin d'estimer leur contribution à l'incidence des symptômes foliaires de l'esca à l'échelle de la parcelle. Le modèle incluait l'identité de la parcelle en tant qu'effets aléatoires indépendants, l'âge de la parcelle modélisée comme un processus autorégressif d'ordre 1 (AR1) et enfin, l'année comme un second processus AR1 spécifique de chaque cépage.

## 3. Résultats

Concernant l'incidence hebdomadaire, les modèles ont indiqué un effet de la sécheresse de l'air et du sol, et de l'évapotranspiration réelle au cours des deux à quatre mois précédant l'expression des symptômes. Plus la transpiration de la plante est élevée (climat humide et chaud), plus l'incidence hebdomadaire de nouveaux cas était élevée. Au contraire, l'incidence estivale de

l'esca diminue avec des printemps chauds et secs (VPD élevé). En outre, nous avons constaté que la température printanière était le principal facteur de la phénologie de la maladie, les symptômes apparaissant plus tôt après des printemps plus chauds. Nos résultats suggèrent qu'un climat plus chaud peut conduire à des symptômes plus précoces de l'esca dans la saison et diminuer son l'incidence en raison de sécheresse accrue (effet direct du réchauffement), puisque l'esca est inhibé en conditions sèches.

Les principaux résultats de l'effet du climat sur l'expression annuelle de l'esca indiquent d'une part, que l'augmentation des températures moyennes pendant la dormance était accompagnée d'une augmentation de l'incidence annuelle de l'esca mais que d'autre part, cette augmentation des températures moyennes avait l'effet inverse pendant les périodes de croissance de la vigne et d'expression des symptômes d'esca. L'augmentation du VPD et la diminution de la disponibilité en eau du sol (SWI) et de l'évapotranspiration du couvert végétal étaient accompagnées d'une diminution de l'incidence annuelle de l'esca. Parmi les différentes périodes phénologiques testées (dormance, période de croissance, période d'expression des symptômes d'esca), toutes présentent des indicateurs écoclimatiques influençant significativement l'incidence de l'esca et particulièrement la période d'expression des symptômes avec la température moyenne et le SWI qui ont les effets les plus fort.

## 4. Conclusion

Nos différentes approches de modélisation ont permis de tester les relations *in situ* entre l'incidence de l'esca (intra- et inter-saisonnière) et le climat, en plus d'autres facteurs de variation tels que le temps, l'âge des parcelles viticoles et le cépage. Nous avons constaté que la corrélation positive intrinsèque entre la température de l'air et son déficit de vapeur d'eau (VPD) était le principal facteur de variation de l'incidence des symptômes foliaires de l'esca au cours de la saison. Des printemps chauds et secs réduisent l'incidence hebdomadaire de l'esca au cours de l'été, avec des symptômes néanmoins plus précoces. De la même manière, la modélisation des données de surveillance nationale va dans le même sens, avec un rôle clé du climat printanier (période de croissance) mais aussi du climat de la période d'expression, de la température et de variables conditionnant la transpiration de la plante. Ces résultats confortent ceux obtenus en conditions contrôlées démontrant le rôle de l'état hydrique de la vigne dans la pathogénèse de l'esca (Bortolami et al. 2021). Cependant, le facteur sous-jacent de ces variations d'incidence pourrait être soit la transpiration de la plante, soit l'activité fongique, ou les deux, ce qui nécessitera de nouvelles études plus approfondies. Enfin, notre étude tend à confirmer que des conditions plus sèches dans le cadre du changement climatique pourraient limiter l'incidence des maladies du bois de la vigne.

# Bibliographie

Allen, C.D., Macalady, A.K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., Kitzberger, T., Rigling, A., Breshears, D.D., Hogg, E.T. and Gonzalez, P. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest ecology and management*, *259*(4), pp.660-684.

Bertsch, C., Ramírez-Suero, M., Magnin-Robert, M., Larignon, P., Chong, J., Abou-Mansour, E., … Fontaine, F. (2013). Grapevine trunk diseases : Complex and still poorly understood. *Plant Pathology*, *62*(2), 243-265. doi: 10.1111/j.1365-3059.2012.02674.x

Bortolami, G., Gambetta, G. A., Cassan, C., Dayer, S., Farolfi, E., Ferrer, N., … Delmas, C. E. L. (2021). Grapevines under drought do not express esca leaf symptoms. *Proceedings of the National Academy of Sciences*, *118*(43). doi: 10.1073/pnas.2112825118

Caubel, J., de Cortázar-Atauri, I.G., Launay, M., de Noblet-Ducoudré, N., Huard, F., Bertuzzi, P. and Graux, A.I., 2015. Broadening the scope for ecoclimatic indicators to assess crop climate suitability according to ecophysiological, technical and quality criteria. Agricultural and forest meteorology, 207, pp.94-106.

Chaloner, T. M., Gurr, S. J., & Bebber, D. P. (2021). Plant pathogen infection risk tracks global crop yields under climate change. *Nature Climate Change*, *11*(8), 710-715. doi: 10.1038/s41558-021-01104-8

Dewasme, C., Mary, S., Darrieutort, G., Roby, J.P. and Gambetta, G.A., 2022. Long-Term Esca Monitoring Reveals Disease Impacts on Fruit Yield and Wine Quality. Plant disease, 106(12), pp.3076-3082.

Fréjaville, T., Vizcaíno-Palomar, N., Fady, B., Kremer, A., & Garzón, M. B. (2020). Range margin populations show high climate adaptation lags in European trees. *Global Change Biology*, *26*(2), 484-495. doi: 10.1111/gcb.14881

Garcia De Cortazar Atauri, I. ; Maury, O. (2019). "GETARI : Generic Evaluation Tool of AgRoclimatic Indicators", DOI 10.15454/IZUFAP, Recherche Data Gouv, V1

Gramaje, D., Urbez-Torres, J.R. and Sosnowski, M.R. (2018). Managing grapevine trunk diseases with respect to etiology and epidemiology: current strategies and future prospects. *Plant disease*, *102*(1), pp.12-39.

Hammond, W.M., Williams, A.P., Abatzoglou, J.T., Adams, H.D., Klein, T., López, R., Sáenz-Romero, C., Hartmann, H., Breshears, D.D. and Allen, C.D. (2022). Global field observations of tree die-off reveal hotter-drought fingerprint for Earth's forests. *Nature Communications*, *13*(1), p.1761.

Lecomte, P., Bénétreau, C., Diarra, B., Meziani, Y., Delmas, C. and Fermaud, M. (2024). Logistic modeling of summer expression of esca symptoms in tolerant and susceptible cultivars in Bordeaux vineyards. *OENO One*, *58*(1).

Le Moigne, P., Besson, F., Martin, E., Boé, J., Boone, A., Decharme, B., Etchevers, P., Faroux, S., Habets, F., Lafaysse, M. and Leroux, D., 2020. The latest improvements with SURFEX v8. 0 of the Safran–Isba–Modcou hydrometeorological model for France. *Geoscientific Model Development*, *13*(9), pp.3925-3946.

Nutter, F.W., Esker, P.D. and Netto, R.A.C. (2006). Disease assessment concepts and the advancements made in improving the accuracy and precision of plant disease data. *European Journal of Plant Pathology*, *115*, pp.95-103.

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P. and Lindgren, F.K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, *4*, pp.395-421.

Torres-Ruiz, J.M., Cochard, H., Delzon, S., Boivin, T., Burlett, R., Cailleret, M., Corso, D., Delmas, C.E., De Caceres, M., Diaz-Espejo, A. and Fernández-Conradi, P., 2024. Plant hydraulics at the heart of plant, crops and ecosystem functions in the face of climate change. *New Phytologist*, *241*(3), pp.984-999.

Vidal, J., Martin, E., Franchistéguy, L., Baillon, M., & Soubeyroux, J. (2010). A 50-year high-resolution atmospheric reanalysis over France with the Safran system. *International Journal of Climatology*, *30*(11), 1627-1644. doi: 10.1002/joc.2003

# Statistique spatiale

# A functional spatial autoregressive model using signatures

Camille Frévent [1]

[1] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France.* camille.frevent@univ-lille.fr

**Résumé.** Nous proposons une nouvelle approche au modèle spatial autoregressif avec covariables fonctionnelles, basé sur la notion de signatures. Celles-ci représentent une fonction comme une série de ses intégrales itérées. Elles présentent l'avantage d'être applicables à un large éventail de processus. Après avoir fourni des garanties théoriques au modèle proposé, nous avons montré dans une étude de simulation que cette nouvelle approche présente des performances compétitives par rapport au modèle traditionnel.

**Mots-clés.** Données fonctionnelles, FSAR, Régression spatiale, Signature, Tenseur

**Abstract.** We propose a new approach to the autoregressive spatial functional model, based on the notion of signature, which represents a function as an infinite series of its iterated integrals. It presents the advantage of being applicable to a wide range of processes.
After having provided theoretical guarantees to the proposed model, we have shown in a simulation study that this new approach presents competitive performances compared to the traditional model.

**Keywords.** Functional data, FSAR, Signature, Spatial regression, Tensor

## 1 Introduction

We are interested here in modelling the relationship between a real-valued random variable $Y$ and a functional covariate $\{X(t), t \in \mathcal{T}\}$ observed in $N$ spatial locations. A traditional approach is to assume that $X$ belongs to $\mathcal{L}^2(\mathcal{T})$, the space of square-integrable functions on $\mathcal{T}$, and to consider the following model:

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \int_{\mathcal{T}} X_i(t)\theta^*(t)\ \mathrm{d}t + \varepsilon_i, \qquad i = 1, \dots, N,\ \ N = 1, 2, \dots$$

where the spatial dependency structure between the spatial units is described by the spatial weights matrix $V_N = (v_{ij,N})_{1 \leq i,j \leq N}$, the autoregressive parameter $\rho^*$ is in a compact space $\mathcal{R}$ and $\theta^* \in \mathcal{L}^2(\mathcal{T})$.

Fermanian (2022) recently investigated the use of signatures in the context of a non-spatial linear regression model with functional covariates. Signatures present the advantages of being applicable to a wide range of processes that are not necessary square-integrable processes.

# 2 The signatures-based spatial autoregressive model

## 2.1 Concept of signatures

Let $\mathcal{T}$ be a compact interval and $X : \mathcal{T} \to \mathbb{R}^p$ be a $p-$dimensional continuous function, $p \geq 2$. Let $(e_i)_{i=1}^p$ be the canonical orthonormal basis of $\mathbb{R}^p$. Then the signature of $X$ can be written as

$$Sig(X) = 1 + \sum_{d=1}^{\infty} \sum_{(i_1,\dots,i_d)} \mathcal{S}_{(i_1,\dots,i_d)}(X) e_{i_1} \otimes \cdots \otimes e_{i_d}. \text{ where } \mathcal{S}_{(i_1,\dots,i_d)}(X) = \int_{\substack{t_1 < \cdots < t_d \\ t_1,\dots,t_d \in \mathcal{T}}} \cdots \int \mathrm{d}X^{(i_1)}(t_1) \dots \mathrm{d}X^{(i_d)}(t_d).$$

## 2.2 Model

We consider the following signatures-based spatial autoregressive model:

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \alpha^* + \langle \theta^*, Sig(X_i) \rangle + \varepsilon_i \tag{1}$$

where the parameter $\theta^*$ is assumed to be written as

$$\theta^* = 1 + \sum_{d=1}^{\infty} \sum_{(i_1,\dots,i_d)} \beta^*_{(i_1,\dots,i_d)} e_{i_1} \otimes \cdots \otimes e_{i_d}.$$

The disturbances $\varepsilon_i$ are assumed to be independent and identically distributed random variables such that $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^{2*}$. They are also independent of $X$.

Then, one can rewrite (1) as

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \alpha^* + 1 + \sum_{d=1}^{\infty} \sum_{(i_1,\dots,i_d)} \beta^*_{(i_1,\dots,i_d)} \mathcal{S}_{(i_1,\dots,i_d)}(X_i) + \varepsilon_i. \tag{2}$$

However, this model cannot be maximized without addressing the difficulty produced by the infinite dimension of the signatures $Sig(X_i)$ (and thus the infinite number of coefficients $\beta^*_{(i_1,\dots,i_d)}$).
Thus we proposed two estimation methods that overcome this challenge.

### 2.2.1 Penalized signatures-based spatial regression

We consider here a slightly modified version of Model (2). We assume that the signature coefficients $\mathcal{S}_{(i_1,\dots,i_d)}(X_i)$ are involved in the model only up to a certain unknown truncation order $D^* \in \mathbb{N}$. The model thus becomes

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \alpha^* + 1 + \sum_{d=1}^{D^*} \sum_{(i_1,\dots,i_d)} \beta^*_{(i_1,\dots,i_d)} \mathcal{S}_{(i_1,\dots,i_d)}(X_i) + \varepsilon_i. \tag{3}$$

Let the signature coefficients vector of $X$, $\mathcal{S}(X)$, be the sequence of all signature coefficients:

$$\mathcal{S}(X) = (1, \mathcal{S}_{(1)}(X), \dots, \mathcal{S}_{(p)}(X), \mathcal{S}_{(1,1)}(X), \mathcal{S}_{(1,2)}(X), \dots, \mathcal{S}_{(i_1,\dots,i_d)}(X), \dots)$$

and the truncated signature coefficients vector at order $D$ of $X$, $\mathcal{S}^D(X)$, be defined as

$$\mathcal{S}^D(X) = (1, \mathcal{S}_{(1)}(X), \mathcal{S}_{(2)}(X), \dots, \underbrace{\mathcal{S}_{(p,\dots,p)}}_{D \text{ terms}}(X)).$$

Then, by noting $s_p(D) = \sum_{d=0}^{D} p^d = \frac{p^{D+1}-1}{p-1}$ the dimension of the truncated signature coefficients vector at order $D$, (3) can be rewritten as

$$S_N \mathbf{Y}_N = \alpha^* \mathbf{1}_N + \xi_{N,D^*} B^* + \boldsymbol{\varepsilon}_N, \tag{4}$$

where $S_N = (I_N - \rho^* V_N)$, $\mathbf{Y}_N$ and $\boldsymbol{\varepsilon}_N$ are two $N \times 1$ vectors of elements $Y_i$ and $\varepsilon_i$, $i = 1, \dots, N$ respectively, $I_N$ denotes the $N \times N$ identity matrix and $\mathbf{1}_N$ denotes the $N \times 1$ vector composed only of 1.

$B_D = (1, \beta_1, \beta_2, \dots, \underbrace{\beta_{(p,\dots,p)}}_{D \text{ terms}})^\top \in \mathbb{R}^{s_p(D)}$, $B^* = (1, \beta^*_1, \beta^*_2, \dots, \underbrace{\beta^*_{(p,\dots,p)}}_{D^* \text{ terms}})^\top \in \mathbb{R}^{s_p(D^*)}$ and $\xi_{N,D}$ is an $N \times s_p(D)$ matrix whose $i^{\text{th}}$ line corresponds to $\mathcal{S}^D(X_i)$.

For a truncation order $D$, the associated conditional quasi log-likelihood function is

$$\begin{aligned} \ell_N(\sigma^2, \rho, \alpha, B_D) &= -\frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) + \ln|S_N(\rho)| \\ &\quad -\frac{1}{2\sigma^2}\left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,D}B_D\right]^\top \left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,D}B_D\right] \end{aligned} \tag{5}$$

where $S_N(\rho) = I_N - \rho V_N$.
Then, Model (4) is estimated using a penalized (ridge) regression.

### 2.2.2 Spatial autoregressive model based on signatures projections

In this section we consider Model (2). Using the notation $B^*_\infty = (1, \beta^*_1, \dots, \beta^*_{(i_1,\dots,i_d)}, \dots)^\top$, and $\xi_{N,\infty}$ the matrix whose $i^{\text{th}}$ line corresponds to $\mathcal{S}(X_i)$, one can rewrite (2) as

$$S_N \mathbf{Y}_N = \alpha^* \mathbf{1}_N + \xi_{N,\infty} B^*_\infty + \boldsymbol{\varepsilon}_N\,, \qquad N = 1, 2, \dots$$

Then, the associated conditional quasi log-likelihood function of the vector $\mathbf{Y}_N$ given $\{Sig(X_i),\ i = 1, \ldots, N\}$ is given by:

$$\ell_N(\sigma^2, \rho, \alpha, B_\infty) = -\frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) + \ln|S_N(\rho)| - \frac{1}{2\sigma^2}\left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,\infty}B_\infty\right]^\top \left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,\infty}B_\infty\right]. \quad (6)$$

**Estimation**
We assume that there exist new coefficients $\boldsymbol{\zeta}_i = (\zeta_{i,1}, \zeta_{i,2}, \ldots)^\top$ and $\Phi^* = (\phi_1^*, \phi_2^*, \ldots)^\top$ such that

$$\langle \theta^*, Sig(X_i)\rangle = \mathcal{S}(X_i)B_\infty^* = 1 + \langle \boldsymbol{\zeta}_i, \Phi^*\rangle,$$

and a positive sequence of integers $C_N$ increasing asymptotically as the sample size $N \to \infty$ such that

$$\langle \boldsymbol{\zeta}_i, \Phi^*\rangle = \sum_{c=1}^{C_N} \zeta_{i,c}\phi_c^* + \sum_{c=C_N+1}^{\infty} \zeta_{i,c}\phi_c^*$$

where the second term vanishes asymptotically when $N \to \infty$.

Then $\langle \theta^*, Sig(X_i)\rangle$ can be approximated by $1+\sum_{c=1}^{C_N} \zeta_{i,c}\phi_c^*$ and $\xi_{N,\infty}B_\infty$ can be approximated by $\mathbf{1}_N + Z_{C_N}\Phi_{C_N}^*$ where $Z_{C_N}$ is the $N \times C_N$ matrix whose $i^{\text{th}}$ line is given by

$$\boldsymbol{\zeta}_i^{C_N\top} = (\zeta_{i,1}, \ldots, \zeta_{i,C_N})$$

and $\Phi_{C_N}^* = (\phi_1^*, \ldots, \phi_{C_N}^*)^\top$.

Then the new parameters can be estimated using a traditional estimation approach for a (non functional) SAR model.

# 3 Finite sample properties

A simulation study was then conducted to compare the performances of the proposed signatures-based spatial autoregressive model considering the penalized spatial regression and the signatures projections strategies. We also compared them with the functional linear model proposed by Ahmed et al (2022).

## 3.1 Design of the simulation study

We considered a grid with $60 \times 60$ locations, where we randomly allocate N=200 spatial units. Then the data was generated according to the following three models where

$$X_i(t) = (X_{i,1}(t), \ldots, X_{i,p}(t))^\top, X_{i,k}(t) = \alpha_{i,k}t + f_{i,k}(t),$$

$$\theta^*(t) = (\theta_1^*(t), \ldots, \theta_p^*(t))^\top, \theta_k^*(t) = \Psi_k t + g_{i,k}(t),$$
$$\alpha_{i,k} \sim \mathcal{U}([-3,3]) \text{ and } \Psi_k \sim \mathcal{U}([-3,3]).$$

$f_{i,k}$ and $g_{i,k}$ are Gaussian processes with exponential covariance matrix with length-scale 1, and $\varepsilon_i \sim \mathcal{N}(0,1)$ for $i \in [\![1, N]\!]$. $X_i$ is observed at 101 equally spaced times of $[0,1]$.

Model 1. $Y_i = \rho^* \sum_{j=1}^N v_{ij,N} Y_j + \int_0^1 X_i(t)^\top \theta^*(t)\, \mathrm{d}t + \varepsilon_i$.
Model 2. $Y_i = \rho^* \sum_{j=1}^N v_{ij,N} Y_j + ||\alpha_i|| + \varepsilon_i$, $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,p})^\top$.
Model 3. $Y_i = \rho^* \sum_{j=1}^N v_{ij,N} Y_j + \langle \mathcal{S}^{D^*}(X_i), \mathcal{S}^{D^*}(\theta^*) \rangle + \varepsilon_i$ where $D^* = 2$.

The spatial weight matrix $V_N$ was constructed using the $k$ nearest neighbors method, and we considered the cases $k = 4$ and $k = 8$, $p = 2, 6, 10$ and $\rho^* = 0, 0.2, 0.4, 0.6, 0.8$.

For each model, several approaches were compared:

1. The approach proposed by Ahmed et al (2022) using a cubic B-splines basis with 12 equally spaced knots to approximate the $X_i$ from the observed data and a functional PCA. As proposed by Ahmed et al (2022), we used a threshold on the number of coefficients such that the cumulative inertia was below 95%. However we also investigated this approach without using a pre-defined threshold on the number of coefficients.

2. Our proposed approach based on the penalized spatial regression.

3. Our proposed approach based on signatures projections. We considered a maximum truncation order for the signatures to reach a maximal number of coefficients of $10^3$. Then a PCA was performed on the truncated signature coefficients vectors. Four strategies were considered: standardizing or not the signature coefficients before computing the PCA and using a threshold on the maximal number of coefficients such that the cumulative inertia was below 95%, or not using a threshold.

For each model and each value of $k, p$ and $\rho^*$, 200 data sets were generated. Each data set was then split into a training, a validation and a test set such that the optimal number of coefficients (for the methods using PCA) or the optimal truncation order (for the penalized regression) was selected on the validation set based on the mean square error (MSE) criterion and the performances were finally evaluated on the test set using the MSE.

## 3.2 Results of the simulation study

All the approaches give similar estimations of the spatial autocorrelation coefficient $\rho^*$ and the latter is reasonably well estimated.

Regarding the MSE, in all cases, our proposed approach based on the PCA on the signature coefficients presents better performances when the signature coefficients are standardized. Not using a threshold on the number of coefficients does not appear to change the performance except for Models 1 and 3 with $p = 10$, where it allows to slightly decrease the MSE. The approach of Ahmed et al (2022) presents similar performances with or without a threshold on the number of coefficients.

With Model 1, which is naturally favourable to the approach of the aforementioned authors, the latter presents the best performances. It should be noted, however, that our approach based on a penalized regression presents close MSEs.

In the case of Model 3 (which is naturally favorable to our methods), our approaches based on a penalized regression or a PCA (with standardization) on the signatures coefficients present much lower MSEs than the approach of Ahmed et al (2022).

Finally, with Model 2, the approach of the above-mentioned authors and our proposition using a PCA (with standardization) on the signatures coefficients give similar performance. Our approach using penalized regression presents slightly lower MSEs.

# 4 Discussion

Here we proposed an alternative to the traditional spatial autoregressive model with functional covariates. This new approach is based on the notion of signatures and presents the advantages of being applicable to a wide range of processes that are not necessary square-integrable processes, and to better capture the differences between the curves. We then proposed two methods for estimating the model, respectively based on a penalized regression and on signatures projections.
The simulation study shows that our approach is competitive with those in the literature.

# Bibliographie

Ahmed, M. S., Broze, L., Dabo-Niang, S., & Gharbi, Z. (2022). Quasi-maximum Likelihood Estimators for Functional Linear Spatial Autoregressive Models. *Geostatistical Functional Data Analysis*, 286-328.

Fermanian, A. (2022). Functional linear regression with truncated signatures. *Journal of Multivariate Analysis*, 192

# Reconstruction géostatistique de la variabilité phénologique spatio-temporelle d'une parcelle viticole

Vu Hoang Ha PHAM [1] & Jean-Pierre DA COSTA [1,2]

[1] *Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, F-33400 Talence, France*
[2] *Bordeaux Sciences Agro, F-33175 Gradignan, France*
*vu-hoang-ha.pham@ims-bordeaux.fr, jean-pierre.dacosta@ims-bordeaux.fr*

**Résumé.** En agriculture de précision, il est important d'appréhender et de gérer la variabilité intra-parcellaire. Différents capteurs permettent de capter temporellement ou spatialement cette variabilité. C'est le cas de capteurs imageurs dont il est facile de tirer des informations sur l'évolution phénologique des cultures : des caméras fixes fournissent des séries temporelles complètes en quelques sites, des caméras embarquées produisent des cartographies de la parcelle entière à quelques dates seulement. Cette étude, conduite sur des données de simulation, a pour objectif général de discuter le gain de connaissance sur la variabilité spatio-temporelle d'une parcelle apporté par l'utilisation conjointe de ces deux types d'information. Nous présentons tout d'abord une méthode de génération de données simulant l'évolution de la surface foliaire en tout point d'une parcelle viticole ainsi que son double échantillonage, spatial et temporel. Nous proposons différentes approches fondées sur des méthodes d'interpolation spatiale ou spatio-temporelle qui tirent partie des deux sources d'information pour reconstruire l'évolution de notre variable sur la parcelle entière et sur l'ensemble de la saison. Une évaluation expérimentale met en évidence la meilleure performance de reconstruction obtenue par une méthode de krigeage spatio-temporel. Elle nous permet également de discuter du mode d'échantillonnage, montrant l'intérêt d'associer quelques séries temporelles ponctuelles à un petit nombre de cartographies spatiales.

**Mots-clés :** données spatiales, séries temporelles, interpolation, krigeage, viticulture

**Abstract.** In precision agriculture, it is important to understand and manage variability within a plot. Various sensors can be used to capture this variability in time or space. This is the case with imaging sensors, from which it is easy to derive information on the phenological evolution of crops: fixed cameras provide complete time series at a few locations, while on-board cameras produce maps of the entire plot at just a few dates. The general objective of this study, carried out on simulation data, is to discuss the gain in knowledge about the spatio-temporal variability of a plot that can be obtained using these two types of information together. First, we present a method for generating data simulating the evolution of leaf area at any point on a vineyard plot, as well as its dual spatial and temporal sampling. We propose different approaches based on spatial or spatio-temporal interpolation methods that take advantage of both sources of information to reconstruct the leaf area evolution over the whole plot and over the whole season. An experimental evaluation highlights the better reconstruction performance obtained by a spatio-temporal kriging method. It also allows us to discuss the sampling method, demonstrating the value of associating a few punctual time series with a small number of spatial mappings.

**Keywords.** Spatial data, time series, interpolation, kriging, viticulture

# 1 Introduction

L'agriculture fait face à différents défis liés notamment au changement climatique et à l'usage de produits phytopharmaceutiques. Pour adapter leurs pratiques, les agriculteurs doivent appréhender la variabilité de leurs parcelles, dans le temps et dans l'espace, afin d'y appliquer les principes d'une agriculture de précision. Cette statégie se base sur la gestion de la variabilité pour améliorer la durabilité de la production agricole (ISPA 2021). Elle recommande l'identification, dans la parcelle étudiée, de zones de gestion aux caractéristiques homogènes (Moral 2010).

Afin d'établir un zonage des parcelles, différentes technologies non invasives sont utilisables. Des capteurs permettent de suivre la croissance et le rendement des cultures tout en caractérisant la variabilité spatiale et temporelle des parcelles. Des solutions de télédétection et de proxidétection se développent partout dans le monde pour surveiller différents systèmes de cultures (Queiroz 2020). En particulier, nous nous intéressons ici aux approches de vision artificielle en viticulture, un des secteurs les plus sensibles à l'application des technologies de l'agriculture de précision (Ammoniaci 2021). L'observation de la vigne présente des difficultés liées à ses caractéristiques culturales, comme une canopée discontinue et une organisation en rangs. Elles ont suscité le développement de capteurs de terrain conçus pour collecter des images au plus près de la culture. Ces capteurs, dits mobiles (e.g. Rançon 2023), sont destinés à être embarqués sur des engins agricoles pour explorer l'ensemble de la parcelle, à raison d'une ou plusieurs images par plante, permettant d'y détecter des maladies (Tardif 2023) ou d'en évaluer le statut azoté (Diago 2015). Au contraire, d'autres capteurs, dits fixes, sont destinés à être installés devant une plante particulière afin d'en observer son évolution temporelle (Rançon 2023).

Chaque type de capteurs présente des avantages et des inconvénients. Un capteur mobile permet de caractériser l'ensemble de la parcelle mais ne peut pas être déployé régulièrement pour des raisons liées aux conditions météorologiques, à l'itinéraire technique ou à la disponibilité des matériels. Les capteurs fixes collectent au contraire des informations journalières sur quelques plantes mais n'assurent pas une couverture suffisante de la parcelle.

Une solution pour tirer parti de ces deux types d'information est d'en effectuer une fusion. La fusion de données peut se définir comme le processus consistant à "combiner les données issues de plusieurs sources pour produire des informations plus cohérentes et plus précises que celles fournies par n'importe quelle source d'une manière individuelle" (Munir 2021). Les méthodes géostatistiques comme le krigeage ou le co-krigeage sont des techniques envisageables pour cette fusion. Elles sont couramment appliquées à des données de proxidétection et de télédétection afin de cartographier des indices de végétation, la conductivité du sol, etc. (Castrignano 2018, Shaddad 2015, Barbedo 2022).

L'objectif général de cette étude est de discuter le gain de connaissances sur la variabilité spatio-temporelle permis par la fusion de ces deux sources d'information : d'une part des données à forte densité spatiale mais faiblement résolues dans le temps et, d'autre part, des données à forte résolution temporelle mais faible résolution spatiale. Nous nous concentrerons sur des informations telles que celles fournies par les capteurs optiques évoqués plus haut, en particulier sur une donnée facile à extraire d'images de vignes comme la surface foliaire.

A ce stade de l'étude, et considérant la difficulté d'instrumenter une parcelle pour en suivre l'évolution sur une saison entière, nous nous limiterons ici à l'utilisation de données simulées.

Plusieurs travaux portent sur les méthodes permettant de simuler les processus physiologiques, dont la surface foliaire , en utilisant des modèles de fonctionnement des cultures comme STICS (Brisson et al. 1998) ou VICTOMO (Mania 2019).Les données simulées par le modèle STICS ont été utilisées pour évaluer les impacts du changement climatique sur le rendement, la phénologie, les conditions de stress ou les besoins en irrigation (Fraga 2018). La variabilité spatiale parcellaire peut aussi être simulée comme une variable auto-corrélée pour étudier les propriétés et les limites des approches d'échantillonnage (Oger 2021).

Inspirés par ces travaux, nous proposons dans cet article une méthode de simulation de données reproduisant l'évolution phénologique d'une parcelle viticole intégrant différentes sources de variabilité spatiale (nature du sol, disponibilité des ressources) et individuelle (matériel végétal), mais également les erreurs de mesure inhérentes à l'utilisation de capteurs imageurs. Par un protocole d'échantillonnage simulant un placement aléatoire de capteurs fixes dans l'espace et d'acquisitions mobiles dans le temps, nous reproduirons différents scénarios d'échantillonnage spatio-temporel de ces données. Nous questionnerons notamment le choix des fréquences d'échantillonnage spatial et fréquentiel. Différentes approches de fusion seront proposées, basées sur des méthodes d'interpolation (inverse-distance) ou des méthodes géostatistiques (krigeage spatial ou spatio-temporel). Enfin, nous évaluerons la capacité de ces approches à estimer la variabilité spatio-temporelle et ainsi à reconstruire l'évolution phénologique de la parcelle.

# 2 Matériels et Méthodes

Cette étude comporte différentes étapes expérimentales représentées à la figure 1 : la simulation de données de surfaces foliaires telles qu'observables par vision embarquée sur une parcelle viticole au cours d'une saison de production ; le sous-échantillonnage spatial ou temporel simulant la prise de données par des capteurs (resp. mobiles ou fixes) ; la reconstruction des données initiales par interpolation spatio-temporelle ; l'évaluation de la reconstruction à l'aide de métriques standards. Chacune de ces étapes est détaillée ci-dessous.

## 2.1 Simulation de données

Le jeu de données simulé est constitué d'une série de cartes journalières exhaustives d'une parcelle d'Avril à Octobre. Le vignoble virtuel comprend 100 rangs de 73 ceps orientés nord-sud. L'inter-rang et la distance entre deux ceps sont fixés à 1,1m. La variable choisie est la surface foliaire dont l'estimation est généralement facile d'accès par vision embarquée.

Afin de simplifier l'étude, nous adoptons quelques hypothèses : i) il n'y a pas de pieds manquants dans la parcelle; ii) les techniques de gestion de canopée (e.g. effeuillage, écimage) ne sont pas appliquées sur la vigne; iii) la parcelle est globalement homogène et ne présente pas de retards phénologique entre ceps ou entre zones de la parcelle.

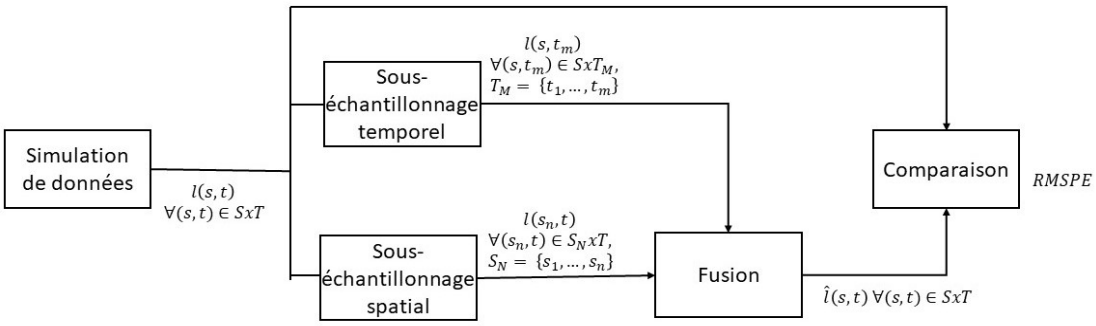



Figure 1: Reconstruction de la variabilité spatio-temporelle par méthodes d'interpolation

On suppose que la variabilité spatio-temporelle de la parcelle est issue de plusieurs composantes : l'évolution phénologique $\varphi(t)$, la variabilité environnementale $g(s)$, la variabilité du matériel végétal $\alpha_p(s)$ et enfin la variabilité de mesure $\alpha_m(s,t)$. La surface foliaire, représentée par la fonction $l(s,t)$ est ainsi donnée par la formule suivante :

$$l(s,t) = \alpha_m(s,t) \times \alpha_p(s) \times g(s) \times \varphi(t) \tag{1}$$

L'évolution phénologique $\varphi(t)$ décrit de manière empirique le développement du feuillage au cours de la saison. Elle est modélisée de manière réaliste par une équation de type sigmoïde :

$$\varphi(t) = \frac{a}{b + \exp\left(-c \times (t-d)\right)}, \forall t \geq 0 \tag{2}$$

où $t$ est le temps (en jours). Les paramètres $a = 0.0623$, $b = 0.0424$, $c = 0.1213$ et $d = 106$ ont été ajustés à partir d'une simulation d'un modèle STICS. L'impact environnemental $g(s)$, seule composante auto-corrélée, décrit la variabilité due aux ressources (i.e. au sol). Elle est générée comme un champ Gaussien aléatoire sans effet de pépite. Le variogramme utilisé, de type exponentiel, est inspiré des données de NDVI collectées sur une parcelle réelle. Les termes relatifs à la variabilité inter-ceps $\alpha_p(s)$ et à l'erreur de mesure $\alpha_m(s,t)$ sont des bruits blancs de lois normales $\mathcal{N}(1,\, \sigma_p^2)$ et $\mathcal{N}(1,\, \sigma_m^2)$. Leurs variances sont choisies pour limiter les variations à respectivement $\pm 10\%$ et $\pm 15\%$. La variabilité inter-cep (liée aux propriétés du matériel végétal lui-même) est considérée comme indépendante du temps mais l'erreur de mesure (liée aux variations de prise d'images) peut varier d'un jour à l'autre.

## 2.2 Echantillonnage

Afin de simuler l'acquisition de données terrain, deux échantillonnages ont été menés. Le premier consiste à collecter $M$ séries temporelles, relatives à $M$ capteurs fixes, échantillonnés aléatoirement sur la parcelle. Afin d'assurer une bonne couverture de la parcelle et éviter toute corrélation spatiale, leurs positions sont tirées selon un processus de Matern II (Baddeley 2016) de sorte que la distance entre deux points soit toujours supérieure à 10 m.

Le second échantillonnage, temporel, consiste à choisir $N$ dates sur la saison, relatives aux $N$ dates d'acquisition de cartes exhaustives par des capteurs mobiles. Ces dates sont choisies aléatoirement lors de la troisième semaine de chaque mois, sauf dans le cas d'une acquisition unique. Dans ce cas, la date est choisie aléatoirement sur tout le mois d'Avril.

## 2.3 Méthodes de reconstruction

Pour chaque scénario d'échantillonnage, un ensemble de $M \in \{1, 3, 6, 15, 24\}$ séries temporelles et $N \in \{1, 2, 4, 7\}$ cartes sont disponibles pour reconstruire l'évolution temporelle de la parcelle tout entière. Pour cela, quatre techniques de reconstruction ont été utilisées qui diffèrent selon la méthode d'interpolation utilisée (interpolation par inverse distance vs. krigeage géostatistique) et selon sa mise en oeuvre (spatiale ou spatio-temporelle).

### 2.3.1 Reconstruction spatiale

Dans ce cas, seuls les $N$ capteurs fixes sont utilisés : l'interpolation est faite à chaque date $t$ indépendamment. La valeur $\hat{l}(s,t)$ en un point inconnu $s$ est estimée comme une moyenne pondérée des valeurs observées aux points $s_i$ à la même date $t$ :

$$\hat{l}(s,t) = \frac{\sum_i w(s_i) l(s_i,t)}{\sum_i w(s_i)}, \tag{3}$$

où les poids $w(s_i)$ diffèrent selon la méthode d'interpolation : IDW ou Krigeage.

**IDW Spatial** – Selon la méthode Inverse Distance Weighting (Burrough 1986), les poids varient selon l'inverse de la distance au point $s$, élevée à une puissance $p$ : $w(s_i) = 1/d^p(s, s_i)$. La valeur de $p$ est fixée à 2 dans nos travaux.

**Krigeage spatial** – La seconde technique utilisée est un krigeage spatial (Cressie 1988). Les poids d'interpolation sont établis par modélisation stochastique de la variabilité spatiale, généralement représentée par le variogramme empirique. A chaque date $t$, la taille $N$ de l'échantillon étant limitée, le variogramme doit être estimé à partir d'une des cartes exhaustives réalisées dans la saison. Nous avons ajusté un modèle de variogramme exponentiel sur les données simulées du mois de juillet, lorsque le pallier de végétation est atteint. Sa portée et son ratio d'anisotropie sont supposés stables dans le temps. En revanche, son effet pépite et son pallier, ont été adaptés pour les stades antérieurs de développement végétatif.

### 2.3.2 Reconstruction spatio-temporelle

Le principe est cette fois-ci d'exploiter à la fois les capteurs fixes et les capteurs mobiles, c'est à dire à la fois les $N$ séries temporelles et les $M$ cartes spatiales :

$$\hat{l}(s,t) = \frac{\sum_{i,j} w(s_i,t_j) l(s_i,t_j)}{\sum_{i,j} w(s_i,t_j)}. \tag{4}$$

Deux approches sont ici aussi utilisées pour le calcul des poids d'interpolation $w(s_i, t_j)$.

**IDW spatio-temporel** – Dérivée de sa version spatiale, l'approche spatio-temporelle traite le temps comme une troisième dimension dans l'expression de la distance. Les dimensions n'étant pas homogènes, il est nécessaire d'appliquer un coefficient de pondération, choisi arbitrairement : $w(s_i, t_j) = 1/d^p(s, t, s_i, t_j)$, avec $d(s, t, s_i, t_j) = d(s, s_i) + \beta \cdot d(t, t_j)$.

**Krigeage spatio-temporel** Les poids de krigeage sont ici déterminés par ajustement d'un modèle exponentiel au variogramme spatio-temporel empirique (Wikle 2019).

La variable d'intérêt (surface foliaire) variant au cours du temps, l'hypothèse de stationnarité spatio-temporelle nécessaire au krigeage spatio-temporel n'est cependant pas garantie. De même, la méthode IDW parait peu pertinente dans le cas d'une non stationnarité temporelle. Pour pallier cette difficulté, il convient d'appliquer ces deux méthodes non pas sur les valeurs brutes mais sur des résidus normalisés :

$$res(s, t) = (l(s, t) - m(t))/m(t) \tag{5}$$

où $m(t)$ décrit l'évolution moyenne de la variable $l$ au cours du temps. En pratique, elle est obtenue en ajustant le modèle de croissance (cf. équation 2) aux données temporelles relatives aux $N$ capteurs fixes.

Enfin, la quantité de données, dans le temps et l'espace, étant très importante, la complexité calculatoire du krigeage spatio-temporel se révèle très élevée. Le krigeage local devient alors une alternative plus avantageuse, qui permet par ailleurs d'assouplir la contrainte de stationnarité spatiale (Graler et al. 2016).

### 2.4 Evaluation

La capacité des méthodes d'interpolation à reconstituer toute l'évolution phénologique de la parcelle est évaluée en comparant, à chaque date $t$, les cartes des surfaces foliaire simulées $\{g(s)\varphi(t), \forall s\}$ (sans variabilité inter-cep ni bruit de mesure) et les résultats d'interpolation $\{\hat{l}(s, t), \forall s\}$. La métrique utilisée est $RMSPE$, racine carrée de l'erreur quadratique moyenne relative, évaluée sur 25 répétitions Monte-Carlo.

## 3 Résultats et Discussion

### 3.1 Jeu de données simulé

Notre méthode de simulation a permis de générer un bloc de données représentant l'évolution de la surface foliaire au cours de la saison en tout pied de vigne de la parcelle. A chaque jour correspond une section du bloc. La troisième dimension décrit le temps. La figure 2 montre un exemple de section, à gauche, sous la forme d'une carte simulant des observations

faites à la date du 10 juillet à l'aide d'un capteur mobile. Elle montre également trois séries temporelles telles qu'observables par des capteurs fixes en trois points de la parcelle.

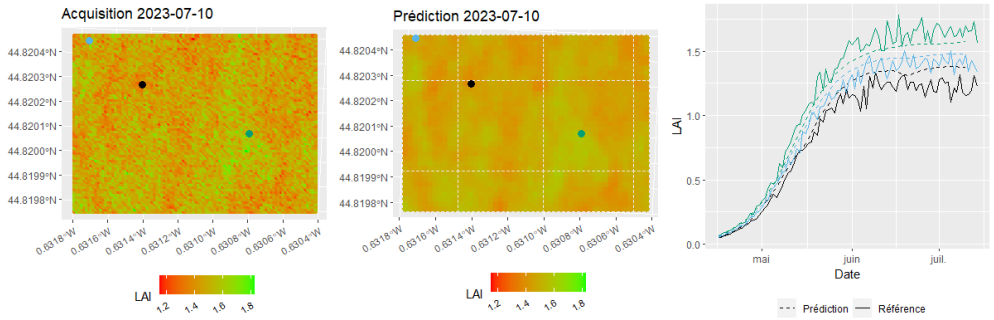



Figure 2: De gauche à droite : carte simulée $l(s, t_j)$ extraite au 10 juillet, résultats d'interpolation $\hat{l}(s, t_j)$ obtenus par krigeage spatio-temporel le même jour (uniquement 1 vigne sur 4), séries temporelles $l(s_i, t)$ et $\hat{l}(s_i, t)$ relatives aux trois points sur la carte.

## 3.2 Performance des méthodes de reconstruction

Les performances des méthodes de reconstruction sont comparées à la figure 3, pour un même nombre de capteurs fixes $N = 3$ et un même nombre de cartes exhaustives $M = 7$.

A un nombre de capteurs fixes limité, il apparait que les méthodes purement spatiales (méthodes (i) et (iii)), qui ne sont basées que sur les $N$ séries temporelles, montrent une performance moindre. La distance entre de capteurs pour assurer une couverture suffisante de la parcelle est trop grande pour détecter les variations spatiales locales conduisant ainsi à des valeurs de RMSPE à la fois plus élevées et plus variables.

La reconstruction par IDW spatio-temporel (méthode (iii)) permet d'exploiter les cartes et les séries temporelles dans l'interpolation. Cette méthode donne des résultats meilleurs et beaucoup plus stables que les méthodes spatiales. Les erreurs sont cependant plus importantes au mois d'Avril, probablement dues aux faibles valeurs de surface foliaire en début de saison. Notons que le coefficient de pondération, fixé ici à $10^{-5}$, a été déterminé en comparant les portées spatiales et temporelles du variogramme spatio-temporel empirique.

Enfin, la reconstruction par krigeage spatio-temporel est celle produisant les erreurs d'interpolation les plus faibles. Une évaluation visualle, cf. figure 2, permet également de constater que cette méthode parvient à restituer la structure spatiale de la parcelle observée sur les données de simulation. C'est cette méthode qui est utilisée pour la suite des expérimentations.

Intéressons nous à présent à la question de l'échantillonnage. Lorsque les acquisitions par capteur mobile se font à raison d'une par mois ($M = 7$), l'ajout de capteurs fixes (au delà de $N = 3$) n'améliore que très peu les résultats de krigeage (Figure 3, en haut à droite). Cela peut être expliqué par le fait que la structure spatiale de la parcelle reste constante au cours

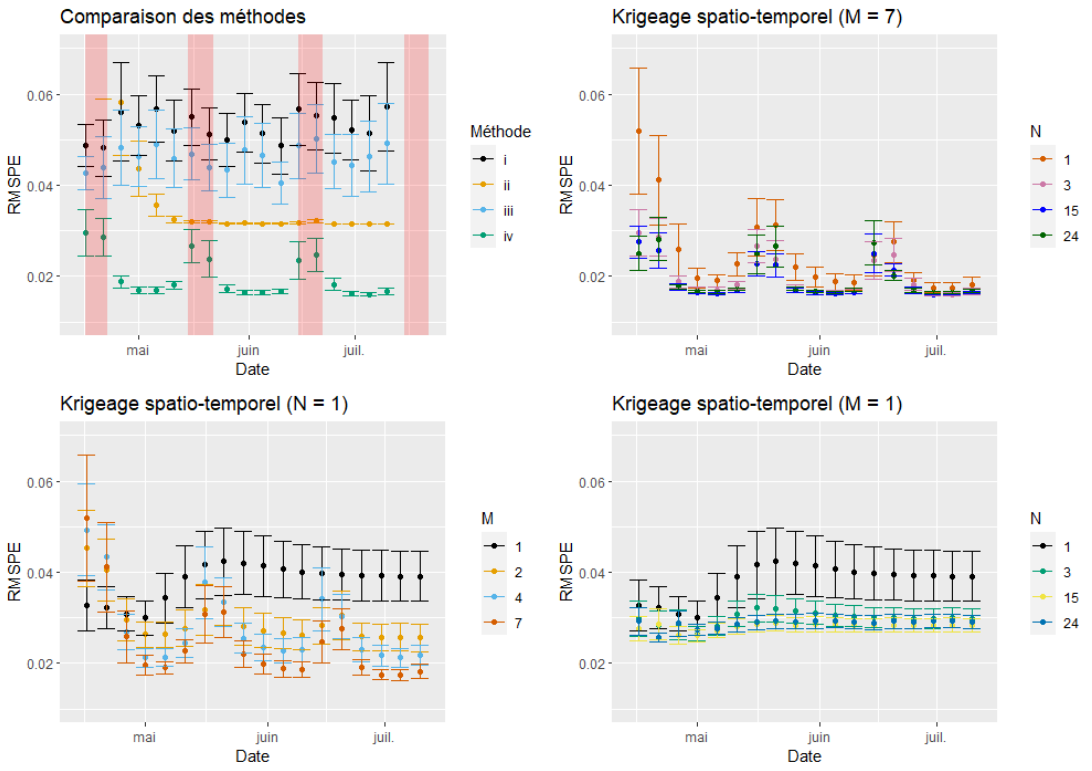



Figure 3: En haut à gauche, comparaison de différentes méthodes de reconstruction avec N = 3 capteurs fixes et M = 7 cartes : (i) IDW spatial, (ii) IDW spatio-temporel, (iii) Krigeage spatial, (iv) Krigeage spatio-temporel). Les rectangles rouges représentant les plages de dates d'acquisition. En haut à droite et ligne du bas : comparaison de différentes stratégies d'échantillonnage dans le cas du krigeage spatio-temporel.

du temps. Observer cette structure à une date et la combiner à l'évolution temporelle obtenue par un petit nombre de capteurs suffit à reconstruire correctement l'ensemble des données. On constate par ailleurs que plus la date à prédire est proche des dates d'acquisitions, plus le RMSPE est important. Ce comportement, qui parait contre intuitif, est probablement dû au fait que les prédictions tendent à se rapprocher des données acquises par les capteurs mais que ces derniers sont bruitées par la variabilité inter-ceps et par l'erreur de mesure.

Lorsqu'un seul capteur fixe est utilisé (cf. Figure 3 en bas à gauche), la diminution du nombre $M$ de cartes exhaustives détériore la qualité de reconstruction. Notons cependant que, dans tous les cas, les performances de reconstruction restent instables et dépendantes de la date à prédire.

Enfin, lorsqu'on se limite à 1 carte exhaustive (Figure 3, en bas à droite), l'ajout de capteurs fixes sur la parcelle permet de mieux reconstruire la variabilité spatiale en diminuant les erreurs d'estimation. Toutefois, l'amélioration s'atténue très vite. Même l'utilisation d'un grand nombre de capteurs ($M = 15$ ou 24) ne permet pas de s'approcher des résultats obtenus avec $M = 4$ ou 7 cartes.

# 4 Conclusion

Dans cet article, nous avons proposé une approche permettant de discuter la stratégie d'échantillonnage d'une parcelle agricole par l'utilisation conjointe de capteurs fixes, fournissant des données temporellement résolues mais spatialement éparses, et de capteurs mobiles, fournissant des données spatialement résolues mais qu'à quelques dates. Cette approche repose sur une méthode de simulation de données de surface foliaire qui intègre à la fois la variabilité des ressources et du matériel végétal, et l'incertitude de mesure. Quatre méthodes ont été implémentées pour la reconstruction de la variabilité spatio-temporelle. Les résultats obtenues sur une parcelle phénologiquement homogène montrent une meilleure efficacité du krigeage spatio-temporel. Cette méthode parvient à reconstruire l'évolution phénologique de la parcelle à partir de quelques séries temporelles et cartes spatialisées.

Dans de futurs travaux, l'étude sera étendue au cas d'une parcelle phénologiquement hétérogène, structurée en plusieurs zones de dynamique phénologique variable. Au delà des métriques classiques, nous chercherons à évaluer la capacité des méthodes à retrouver ces différentes zones. Nous envisageons également d'appliquer cette approche à des données réelles acquises sur une parcelle viticole à l'aide de capteurs imageurs, fixes et embarqués.

# Bibliographie


Ammoniaci, M., Kartsiotis, S.-P., Perria, R. et Storchi, P. (2021), State of the Art of Monitoring Technologies and Data Processing for Precision Viticulture. *Agriculture*, 11(3), p. 201.

Baddeley, A., Rubak, E. et Turner, R. (2015), *Spatial Point Patterns: Methodology and Applications with R*. London : Chapman and Hall/CRC Press.

Barbedo, J. G. A. (2022), Data Fusion in Agriculture: Resolving Ambiguities and Closing Data Gaps. *Sensors*, 22 (6), pp. 2285.

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M.-H., Ruget, F., Nicoullaud, B., Gate, Ph., Devienne-Barret, F., Antonioletti, R. et Dürr, C. (1998), STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. *Agronomie*, 18, pp. 311-346.

Burrough, P. A. (1986), Principles of Geographical Information systems for land resource assessment. *Oxford University Press, Oxford*.

Castrignano, A., Buttafuoco, G., Quarto, R., Parisi, D., Viscarra Rossel, R.A., Terribile, F., Langella, G. et Venezia, A. (2018), A geostatistical sensor data fusion approach for delineating homogeneous management zones in Precision Agriculture. *CATENA*, 167, pp. 293-304.

Cressie, N. (1988), Spatial prediction and ordinary kriging. *Mathematical Geology*, 20 (4), pp. 405-421.

Diago, M. P. et Tardaguila, J. (2015), Vinerobot: On-the-go vineyard monitoring with non-invasive sensors. *Progres Agricole et Viticole*, pp. 1-4.

Fraga, H., García de Cortázar Atauri, I. et Santos, J.A. (2018), Viticultural irrigation demands under climate change scenarios in Portugal. *Agricultural Water Management*, 196, pp. 66-74.

Gräler, B., Pebesma, E. J. et Heuvelink, G. B. M. (2016), Spatio-Temporal Interpolation using gstat. *R Journal*, 8, pp. 204.

ISPA, International Society for Precision Agriculture (2021), Precision Ag Definition. En ligne : https://www.ispag.org/about/definition [Consulté le 4 octobre 2023].

Mania, E., Andreoli, V., Cavalletto, S., Cassardo, C. et Guidoni, S. (2019), VICMOTO: Physical modeling and numerical simulation applied to vineyard. Poni, S. (éd.), *BIO Web of Conferences*, 13, pp. 02006.

Moral, F. J., Terrón, J. M. et Marques da Silva, J.R. (2010), Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. *Soil and Tillage Research*, 106 (2), pp. 335-343.

Munir, A., Blasch, E., Kwon, J., Kong, J. et Aved, A. (2021), Artificial Intelligence and Data Fusion at the Edge. *IEEE Aerospace and Electronic Systems Magazine*, 36 (7), pp. 62-78.

Oger, B., Vismara, P. et Tisseyre, B. (2021), Combining target sampling with within field route-optimization to optimise on field yield estimation in viticulture. *Precision Agriculture*, 22 (2), pp. 432-451.

Queiroz, D. M., Coelho, A. L. F., Valente, D. S. M. et Schueller, J. K. (2020), Sensors Applied to Digital Agriculture: A Review. *Revista Ciência Agronômica* [en ligne], 51 (5).

Rançon, F., Keresztes, B., Fontaine, G., Deshayes, A., Da Costa, J.-P. et Germain, C. (2023), Vinelapse: an autonomous grapevine observation image sensor. In: *ECPA 2023*, Bologna, Italy.

Rançon, F., Keresztes, B., Deshayes, A., Tardif, M., Abdelghafour, F., Fontaine, G., Da Costa, J.-P. et Germain, C. (2023), Designing a Proximal Sensing Camera Acquisition System for Vineyard Applications: Results and Feedback on 8 Years of Experiments. *Sensors*, 23 (2), pp. 847.

Shaddad, S. M., Madrau, S., Castrignano, A. et Mouazen, A. M. (2016), Data fusion techniques for delineation of site-specific management zones in a field in UK. *Precision Agriculture*, 17 (2), pp. 200-217.

Tardif, M., Amri, A., Keresztes, B., Deshayes, A., Martin, D., Greven, M. et Da Costa, J.-P. (2022), Two-stage automatic diagnosis of Flavescence Dorée based on proximal imaging and artificial intelligence: a multi-year and multi-variety experimental study. *OENO One*, 56 (3), pp. 371-384.

Wikle, C.K., Zammit-Mangion, A. et Cressie, N., 2019. *Spatio-Temporal Statistics with R.* Chapman&Hall/CRC, Boca Raton, FL.

# DETECTION OF RESIDUAL BLOCKS IN GRID-BASED DATA USING TREE SEGMENTATION

Karen Wolf [123] & Pierre Fernique [2] & Hans-Peter Piepho [3]

[1] *Limagrain Europe, Germany karen.wolf@limagrain.com*
[2] *Limagrain Europe, France pierre.fernique@limagrain.com*
[3] *University of Hohenheim, Germany hans-peter.piepho@uni-hohenheim.de*

**Abstract.** Applying spatial models to grid-based data relies on strong hypotheses, notably, stationarity. We propose a method to explore the hypothesis of stationarity and pinpoint parts of the grid-based data, where the hypothesis is violated.

We apply these methods to field trials in plant breeding, where grid-based data corresponds to measurements and coordinates of plant varieties in the field. The hypothesis of stationarity may be violated if spatial patterns exist in the residuals in certain areas of the field (i.e., residual blocks), due to environmental effects.

Our methods represent grid-based data in a tree graph using quad and binary tree. Residual blocks are then recovered using tree segmentation of the tree indexed data.

**Keywords.** Grid-based data, Spatial models, Gaussian Random Fields, Stationarity, Tree indexation, Tree segmentation, Plant breeding

# 1 Introduction

**Plant Breeding** To ensure food security despite global challenges such as climate change, steadily developing plant diseases, and increasing human population, the continuous improvement of plant varieties (i.e., cultivars) is crucial. Thereby, efficient plant breeding plays an essential role in achieving this goal. It aims at producing new cultivars with enhanced traits such as abiotic resistances (e.g., drought tolerance), biotic resistances (e.g., insects), and higher yield (Qaim, 2020).

**Field Trials** Plant breeders select cultivars based on their performance. This performance is measured in field trials, where the cultivar effect is computed to represent this performance. Field trials are in-vivo experiments. They are thus carried out under the influence of macro-environmental effects (e.g., rain or solar radiation affecting cultivar effect). Cultivars are therefore tested in and selected for specific macro-environments (i.e., geographic zones with homogeneous environmental conditions) where they can be cultivated by farmers.

The particularities of field trials in plant breeding are described by Mackay et al. (2019). Notably, plants cannot move. They are cultivated within plots which are usually small rectangular areas within the field, where they remain during the entire cultivation period. Therefore, in addition to macro-environmental effects, the cultivar effect of each plant is strongly affected by micro-environmental effects (e.g., variation in soil quality, shadowing of neighbouring plants). All these

undesirable micro-environmental effects are regrouped together into the generic field effect term. In the conduction of field trials, it is of high priority to estimate cultivar effects accurately and remove all micro-environmental effects from this estimation.

**Experimental Designs** One major problem when trying to conduct field trials to get an indication of the cultivar effect is that it may be confounded with the field effect. If this is the case, cultivar effect estimation is inaccurate and so are breeding decisions.

To overcome this problem, breeders use experimental designs. By specific spatial allocations of replications and randomisations of the cultivars, they can take into account *a priori* information relative to field conditions (e.g., soil uniformity, slopes, irrigations). For example, breeders can specify so-called blocks, which are areas in the field within which irrigation conditions are assumed to be homogeneous (Gomez and Gomez, 1984). The specifications of an experimental design can be represented in a design matrix $\boldsymbol{X}$ which encodes the cultivar effect of interest and several other effects such as various block effects to capture the field effect.

**Trial Analysis** During, or at the end, of the cultivation period, breeders collect data from each plot. In the case of metric traits such as yield, the data can be assumed to follow a Gaussian Random Field (GRF), denoted as follows,

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of effects, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix. $\boldsymbol{X}$ can be either given directly by the experimental design or altered to remove and add block effects *a posteriori*. Different types of GRFs exist, for example $IID$ (i.e., $\boldsymbol{\Sigma} = \sigma^2\mathbb{I}$) or $AR1$ (e.g., $\boldsymbol{\Sigma} = \sigma^2 \cdot \left(\boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c\right)$, for more details see Butler et al. (2017)).

In general, the combinations of all possible design matrices $\boldsymbol{X}$ and all possible variance-covariance matrices $\boldsymbol{\Sigma}$ are used to set up a collection of models. And, among these models, the most appropriate modeling of the data can be found using model selection based on BIC, AIC, or cross-validation. Using the most appropriate model, we can restore the cultivar effect and use it for further analyses.

**Residual Blocks** Using the processes described hereabove, breeders assume several hypotheses, one of the strongest being stationarity. In this case, stationarity implies constant expectation (i.e., $\boldsymbol{\beta}$) and constant variance (i.e., $\boldsymbol{\Sigma}$) across the whole trial.

Stationarity can be assumed when all field conditions are known and included in the experimental design. However, not all these conditions are necessarily known and considered *a priori* in the experimental design. We here focus on the later case, that could lead to the violation of the assumption of stationarity. For example, if the field trial is very large, not all plots can be harvested on the same day. If this is not taken into account in the experimental design by including a block effect, cultivar effect estimation will be inaccurate. Furthermore, variance cannot always be assumed to be constant across the whole field trial, especially for large trials. Indeed, soil conditions may be locally homogeneous, but globally very heterogeneous, leading to large differences in terms of variance, notably for distant areas.

If the hypothesis of stationarity is violated, spatial patterns remain in the field trial. Residuals in different parts of the field follow a distinct GRF within each part. We will refer to the affected areas as *residual blocks*. Note that, if resources were unlimited, an appropriate number of replications could compensate for an imprecise experimental design. However, in practice, the possible number of replications is limited which leads to inaccurate cultivar effect estimation. Therefore, to avoid inaccurate breeding decisions, plant breeders would need to discard field trials where the hypothesis of stationarity is violated. However, if inaccuracy is induced by detectable residual blocks, plant breeders would not need to discard the whole trial. Instead, they could discard only those plots where the hypothesis of stationarity is violated.

In some cases, visualising the residuals of the estimated model in a heatmap may reveal those residual blocks. However, these manual investigations of the stationarity hypothesis are not always easy and methods able to screen a large number of trials to detect residual blocks could benefit plant breeding.

# 2 Detection of Residual Blocks

To detect residual blocks, we use tree indexation to get a multi-scale representation of grid-based data (i.e., matrix) and tree segmentation to find the best representative scale of this matrix.

## 2.1 Definition of a Tree Graph [1]

In graph theory, a graph $G$ is defined by two sets, vertices $V \subset \mathbb{N}$ and edges $E$, with

$$\emptyset \subseteq E \subseteq \left\{(u, v) \in V^2 \mid u \neq v\right\}.$$

Some definitions are needed to apply the concept of tree graphs to the detection of residual blocks:

**child set** of vertex $v$, denoted as $ch(v)$, is defined as follows:
$\forall v \in V, ch(v) = \{u \in V \mid (v, u) \in E\}$.

**descendant set** of vertex $v$, denoted as $de(v)$, is defined as follows:
$\forall v \in V, de(v) = \left\{\cup_{u \in ch(v)} de(u)\right\} \cup ch(v)$.

**parent set** of vertex $v$, denoted as $pa(v)$, is defined as follows:
$\forall v \in V, pa(v) = \{u \in V \mid (u, v) \in E\}$.

**leaf set** of vertex $v$ denoted as $le(v)$, or a set of vertices $A$, is defined as follows:
$\forall v \in V, le(v) = \left\{u \in de(v) \mid ch(u) = \emptyset\right\}$,
$\forall A \subseteq V, le(A) = \cup_{v \in A} le(v)$.

Since we use tree graphs to index grid-based spatial data, we associate a set of coordinates to each vertex in $V$. Therefore, instead of indices, we use coordinates to represent vertices, hence,

$$V \subset P\left(\{(i, j) \in [0\,..\,R[ \times [0\,..\,C[\}\right),$$

[1] The definitions used are mostly based on Lauritzen (1996).

where $R$ is the number of rows and $C$ is the number of columns of the matrix and $\mathcal{P}(\cdot)$ denotes the power set. To facilitate manipulation of coordinates, let $I_v$ (resp. $J_v$) be the set of rows (resp. columns) of a vertex $v$:

$$\forall v \in V, \qquad I_v = \{i \in [0\,..\,R[\mid \exists j \in [0\,..\,C[\wedge(i,j) \in v\}\,,$$
$$J_v = \{j \in [0\,..\,C[\mid \exists i \in [0\,..\,R[\wedge(i,j) \in v\}\,.$$

## 2.2 Tree indexation

Using the terms of Hunter and Steiglitz (1979), a two-dimensional array of information (i.e., matrix) can be represented by a tree graph. This can be done using tree indexation. To cover the full information of the matrix, the tree graph must have as many leaves as the matrix has elements.

The leaves are the greatest depth of the tree and the least compact representation of the matrix. All smaller depths represent the matrix more compactly, with the smallest depth, consisting of one single vertex, as its most compact representation. Thus, the tree can be regarded as a multi-scale representation of the matrix with the greatest depth being the finest scale of the matrix, and the smallest depth being the coarsest scale of the matrix.

All vertices are associated with a subset of the matrix. In our case, children are a nested split of the matrix of their parent. The connection of parents to their children (i.e., matrices and their nested split) is represented by the edges of the tree.

Applied to the example of field trials, a vertex is a block of the field. Its children are one possible division of this block and leaves are plots.

To construct a tree indexation of a matrix, we use quad and binary tree algorithms. In every step $t \in \mathbb{N}$ of the algorithm, the tree is updated. At $t = 0$, the tree $T_0$ consists of a single vertex $v_0$, which covers the information of the whole matrix:

$$T_0 = (V_0, E_0) \text{ with } V_0 = \{v_0\}, \qquad E_0 = \emptyset,$$
$$v_0 = \{(i,j) \in [0\,..\,R[\times[0\,..\,C[\,\}\,.$$

**Deterministic quad tree** For quad trees, vertices are either leaves or have four children. In our case, the quad tree divides a vertex into its children in a deterministic manner. Thus, the matrix is subdivided into four equal submatrices along the rows or along the columns:

$$\forall t \in \mathbb{N}, \qquad V_{t+1} = V_t \underset{v \in le(V_t)}{\cup} ch(v),$$
$$ch(v) = \left\{v_{\boxed{1}}, v_{\boxed{2}}, v_{\boxed{3}}, v_{\boxed{4}}\right\},$$

with $v_{\boxed{1}} = \left\{(i,j) \in v \mid i < \lceil m(I_v)\rceil, j < \lceil m(J_v)\rceil\right\}$, $v_{\boxed{2}} = \left\{(i,j) \in v \mid i < \lceil m(I_v)\rceil, j \geq \lceil m(J_v)\rceil\right\}$,
$v_{\boxed{3}} = \left\{(i,j) \in v \mid i \geq \lceil m(I_v)\rceil, j < \lceil m(J_v)\rceil\right\}$, $v_{\boxed{4}} = \left\{(i,j) \in v \mid i \geq \lceil m(I_v)\rceil, j \geq \lceil m(J_v)\rceil\right\}$,

where $m(\bullet) = \frac{max(\bullet)-min(\bullet)}{2} + min(\bullet)$.
Furthermore, the tree is updated by the corresponding directed edges,

$$\forall t \in \mathbb{N}, \qquad E_{t+1} = E_t \underset{v \in le(V_t)}{\cup} \left\{(v, v_{\boxplus}), (v, v_{\boxplus}), (v, v_{\boxplus}), (v, v_{\boxplus})\right\}.$$

**ML binary tree** For binary trees, vertices are either leaves or have two children. We use a binary tree that divides a vertex into its two children (i.e., a matrix into two submatrices, along the rows or along the columns) based on Maximum Likelihood (ML). We define

$$\begin{aligned} \forall t \in \mathbb{N}, \forall v \in le(V_t), \qquad \forall x \in I_v, \qquad & v^x_{\blacksquare} = \{(i,j) \in v \mid j < x\}, \\ & v^x_{\blacksquare} = \{(i,j) \in v \mid j \geq x\} = \overline{v^x_{\blacksquare}}, \\ \forall y \in J_v, \qquad & v^y_{\blacksquare} = \{(i,j) \in v \mid i < y\}, \\ & v^y_{\blacksquare} = \{(i,j) \in v \mid i \geq y\} = \overline{v^y_{\blacksquare}}. \end{aligned}$$

For a possible division of a vertex $v$ into its children $\{v^x_{\blacksquare}, \overline{v^x_{\blacksquare}}\}$ (or $\{v^y_{\blacksquare}, \overline{v^y_{\blacksquare}}\}$) we can fit a separate spatial model to each of the induced submatrices, as described in Section 1 for the modeling of the data of field trials. For each possible division, we can therefore compute the log-likelihood and among all possible divisions of each leaf at step $t \in \mathbb{N}$, we select the one which maximises the log-likelihood. We will refer to the best possible division as $\{v^*, \overline{v^*}\}$ in the following.

The vertices $\{v^*, \overline{v^*}\}$ and their corresponding edges $\{(v, v^*), (v, \overline{v^*})\}$ are used to update the set of vertices $V_t$ and edges $E_t$ of the tree $T_t$:

$$\begin{aligned} \forall t \in \mathbb{N}, \qquad & V_t \underset{v \in le(V_t)}{\cup} \left\{(v^*, \overline{v^*})\right\}, \\ & E_t \underset{v \in le(V_t)}{\cup} \left\{(v, v^*), (v, \overline{v^*})\right\}. \end{aligned}$$

For both, quad tree and binary tree, tree indexation stops at $t_{max}$, where the tree consists of as many leaves as the matrix has elements. This implies that there is no leaf $v$ of $T_{t_{max}}$ that still has children. $T_{t_{max}} = \left(V_{t_{max}}, E_{t_{max}}\right)$ will be referred to as $T = (V, E)$ in the following.

## 2.3 Tree Segmentation

As described in the previous section, a multi-scale representation of a matrix can be obtained by tree indexation, resulting in the tree $T = (V, E)$. All vertices of the tree are associated with a subset of the matrix. The set of vertices of the greatest depth (i.e., the leaves) represents the finest scale of the matrix, whereas the set of vertices of the smallest depth (i.e., $v_0$) represents the coarsest scale of the matrix.

If residuals can be separated into different parts, each following a distinct GRF, residual blocks correspond to a representation of the matrix in between the extremes of the finest and the coarsest scale. They can thus be represented by a subset of vertices of the tree, referred to as change-point set. Note that, vertices of the change-point set may be part of a different depth.

To estimate the change-point set (i.e., the best scale of the matrix) we use tree segmentation.

Let $\mathcal{P}(V)$ be the powerset of $V$. We define $\mathcal{P}'(V)$ as a subset of $P(V)$ such as

$$\mathcal{P}'(V) = \{P \in \mathcal{P}(V) \setminus \emptyset \mid le(P) = le(V)\}. \tag{1}$$

In the following, we will work with $\mathcal{P}''(V) \subset \mathcal{P}'(V)$, where additionally

$$\mathcal{P}''(V) = \left\{P \in \mathcal{P}'(V) \mid \forall (p, q) \in P^2,\ le(p) \cap le(q) = \emptyset\right\}. \tag{2}$$

The leaves of each vertex $v$ of a possible change-point set, $P \in \mathcal{P}''(V)$, cover a fraction of the matrix. We can fit a separate spatial model to each of the induced submatrices of the division of a vertex to its children. Thus, for each possible change-point set $P \in \mathcal{P}''(V)$, we can calculate the log-likelihood.

Maximising the log-likelihood over all possible change-point sets will lead to overfitting since it results in choosing the most complex model, where each leaf of the tree represents one residual block (i.e., $P = le(T)$). Thus, we are using a score to account for the trade-off between bias (i.e., too few residual blocks) and overfitting (i.e., too many residual blocks).

Estimating the optimal change-point set within all possible change-point sets (i.e. $P \in P''(V)$), is a combinatory problem. Therefore, we use a heuristic approach, where we recover the residual blocks covered by $P$ iteratively based on tree segmentation using two algorithms, Top-Down and Bottom-Up.

The Bottom-Up algorithm iteratively recovers residual blocks from the greatest depth of the tree (i.e., leaves) to the smallest depth (i.e., $v_0$). Based on a score, it decides if merges of vertices to their parents are accepted. The Top-Down algorithm, iteratively recovering residual blocks from $v_0$ to the leaves, decides if a split of a vertex to its children is accepted. As a score, we use BIC.

# 3 Results and Discussion

**Material and Data** To test the performance of the tree segmentation algorithms, we conduct a simulation study. We simulate field trials based on the geometry (i.e., the number of rows and columns and the coordinates for missing data) of already conducted field trials from historical Limagrain data. This ensures a realistic geometry of our simulations.

The different steps of the simulation are described in Figure 1. They rely on the simulation of blocks using Lloyd algorithm (Lloyd, 1982). Each parameter $\theta$ of a GRF is simulated using a zero-inflated model (see Eggers (2015)). The parameters are either 0, with a probability of 0.5, or drawn from different distributions:

- $\rho_r$, $\rho_c$, and $\beta$ are drawn from a Beta distribution, with $\mathcal{B}(\alpha, \beta)$, with $\alpha = 3$ and $\beta = 2$.
- $\sigma$ is drawn from a Gamma distribution, $\mathcal{G}(k, \theta)$, with $k = 3$ and $\theta = 1/3$.
- $\phi$ is drawn from an Exponential distribution, $\exp(1/\lambda)$, with $\lambda = 0.25$.

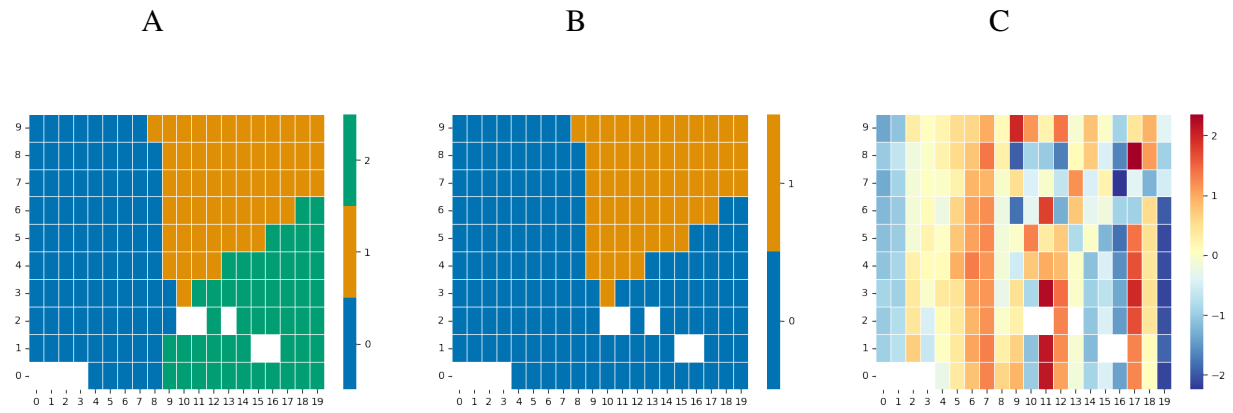



**Figure 1:** Illustration of the simulation of residual blocks. This simulation is based on the geometry of a field trial with 10 rows and 20 columns and missing data. Three blocks are simulated using Lloyd algorithm based on euclidean distance (A). Two simulated blocks are grouped into one resulting in two blocks (B), leading to more convex and concave shapes of the blocks than a direct simulation of two blocks. Residuals within each block follow a distinct simulated GRF (C).

**Performance of tree segmentation** We compare the recovered residual blocks of the tree segmentation algorithms with the true residual blocks of the simulations. To get an indication of how well recovered and true residual blocks coincide, we calculate a measure of similarity (Figure 2). The measure of similarity is based on the Hungarian algorithm (see Kuhn (1955), Denœud and Guénoche (2006)), either without clustering or by assuming perfect clustering (oracle clustering). The ML binary tree is a data-driven construction from top to bottom. We therefore expect meaningful residual blocks at small tree depth. Conversely, the quad tree is a deterministic construction. We therefore expect that segmentation is meaningful at great tree depths and that only clustering of this segmentation could lead to meaningful residual blocks.

For quad tree, Bottom-Up recovers more residual blocks than Top-Down (see Figure 2, E). The Bottom-Up algorithm tends to recover more residual blocks (i.e., vertices of $P$ being part of greater depths of the tree) than the Top-Down algorithm since it recovers residual blocks from the greatest depth of the tree to the smallest depth. A larger number of recovered residual blocks has a higher chance of being clustered to the true residual blocks by oracle clustering. Therefore, with oracle clustering, Bottom-Up shows higher similarities than Top-Down (see Figure 2, A and B).

For binary tree, both algorithms recover a similar number of residual blocks (see Figure 2, E). No major difference in similarities can be found between the two algorithms (see Figure 2, C and D).

The Bottom-Up algorithm recovers more residual blocks for quad tree than for binary tree (see Figure 2, E). As mentioned earlier, separations of the binary tree at greater depths are less meaningful than for quad tree. Thus, Bottom-Up algorithm accepts fewer merges of vertices to their parent (equivalent to accepting more splits) for quad tree and stops at a greater depth of the tree (i.e., recovers more residual blocks).

The Top-Down algorithm recovers more residual blocks for binary tree than for quad tree (see

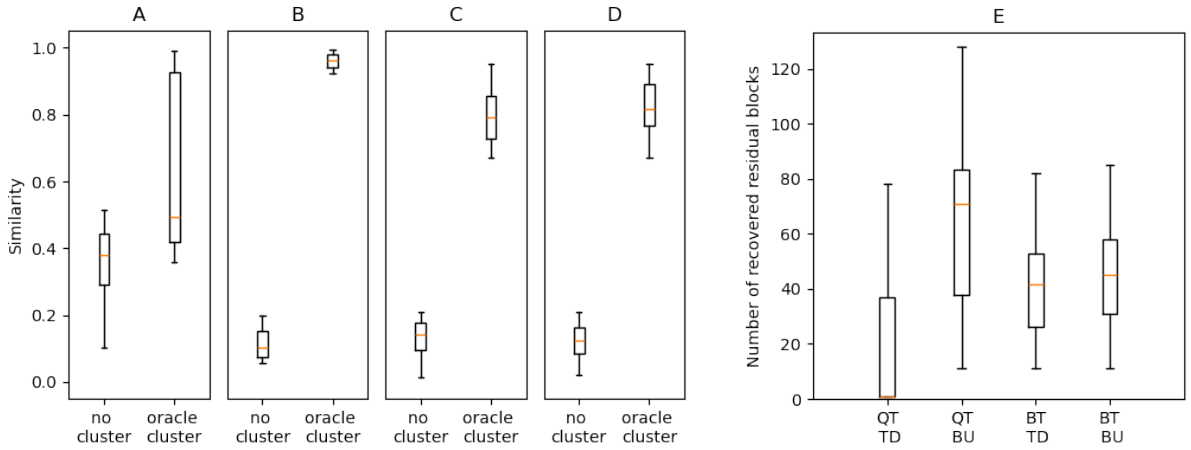



**Figure 2:** Boxplots without outliers for similarities between recovered and true residual blocks. A (resp. B, C, D) Similarity of quad tree with Top-Down algorithm (resp. quad tree with Bottom-Up algorithm, binary tree with Top-Down algorithm, binary tree with Bottom-Up algorithm). Similarities for recovered residual blocks without clustering are calculated using the Hungarian algorithm. Similarities for recovered residual blocks with oracle clustering are calculated using a recursive Hungarian algorithm.
(E) Number of recovered blocks of quad tree and binary tree with Top-Down and Bottom-Up algorithm.

Figure 2, E). Separations of a vertex into two (i.e., binary tree) have a higher chance of being accepted than separations into four (i.e., quad tree). Therefore, in comparison to quad tree, the Top-Down algorithm accepts splits of vertices until greater depths for binary tree (i.e., recovers more residual blocks).

For quad tree, where we aim at recovering more residual blocks, the Bottom-Up algorithm should be used. For binary tree, where we aim at recovering less residual blocks, the Top-Down algorithm is more adequate.

**Constraints for tree segmentation** Estimating $P$ to recover residual blocks is a model selection problem where we are not looking for optimal modeling of only a fraction of the trial (i.e., one vertex) but of the whole trial (i.e., tree graph). For the estimation of $P$, we use constraint (2) to ensure that the information covered by each vertex of the change-point set, is not covered by any other vertex of the change-point set. Otherwise, the segmentation of one part of the tree would depend on other segmentations and more possible solutions for $P$ would exist. This is the case in Fernique (2014), where the tree graph has a biological meaning since it represents a real tree. Thus, for segmentation, the whole structure of the tree graph (i.e., real tree) must be accounted and is of particular interest. Since we use the tree graph only as a tool to recover residual blocks, without biological meaning, the tree structure is of less interest and a less complex set of change-points can be considered, allowing for more computational efficiency.

**Clustering of tree segmentation** Constraint (2) also implies that residuals within several recovered residual blocks may follow the same GRF. As in Fernique (2014), a clustering algorithm must

account for the constraint that all elements of one recovered residual block are clustered with all elements of other recovered residual blocks. This can be done using a constrained Viterbi EM algorithm (Viterbi, 1967). Hereabove, we assume perfect clustering of recovered residual blocks to the true residual blocks (i.e., oracle clustering). Therefore, similarities are always higher with than without clustering (see Figure 2, A to D).

For quad tree, the Top-Down algorithm stops already at small depths of the tree (i.e., recovers low numbers of residual blocks) in comparison to the Bottom-Up algorithm with quad tree and in comparison to both algorithms with binary tree (see Figure 2, E). As already mentioned, this is due to the fact that separations into more parts (e.g., into four for quad tree) have a lower chance of being accepted than separations into fewer parts (e.g., into two for binary tree).

Therefore, especially for Top-Down algorithm with quad tree, synchronous segmentation-clustering might lead to higher similarities. The algorithm could test if splits of vertices into one, two, three, or four are accepted or not, leading to a higher chance of acceptance of splits in general and thus to more recovered residual blocks. Synchronous segmentation-clustering is less applicable for the Bottom-Up algorithm since it starts with a high number of residual blocks.

**Scores for tree segmentation** As score for tree segmentation, we use BIC. However, if tree segmentation leads to oversegmentation (i.e., recovers too many residual blocks), a subsequent clustering algorithm might not find the real residual blocks. Especially, the Bottom-Up algorithm with quad tree recovers a large number of residual blocks (see Figure 2, E). With oracle clustering it results in higher similarities than Top-Down with quad tree but without oracle clustering it results in lower similarities than Top-Down with quad tree. Thus, oversegmentation should be avoided (see Figure 2, A and B).

For segmentation, BIC was shown to tend to oversegmentation since its penalisation is too liberal (Zhang and Siegmund, 2007). Instead, other penalisations, such as slope heuristics could be used (Birgé and Massart, 2007).

**Application of tree segmentation in plant breeding** As Gilmour (2000) argues, '[post-blocking] based purely on statistical significance of arbitrary contrasts without a plausible explanation is not justified.' Our methods are meant to support plant breeders in screening large numbers of field trials to give them indications about areas in the trials that induce inaccuracy to their cultivar effect estimations. Based on their observations in the field during the cultivation period and data collection, plant breeders must decide if they want to discard those plots where the hypothesis of stationarity is violated.

# References


Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138:33–73, 2007.

DG Butler, BR Cullis, AR Gilmour, BJ Gogel, and R Thompson. Asreml-r reference manual version 4. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK*, 2017.

Lucile Denœud and Alain Guénoche. Comparison of distance indices between partitions. In *Data Science and Classification*, pages 21–28. Springer, 2006.

Julia Eggers. On statistical methods for zero-inflated models, 2015.

Pierre Fernique. *A statistical modeling framework for analyzing tree-indexed data: Application to plant development on microscopic and macroscopic scales*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc, 2014.

Arthur R Gilmour. Post blocking gone too far! recovery of information and spatial analysis in field experiments. *Biometrics*, 56(3):944–945, 2000.

Kwanchai A Gomez and Arturo A Gomez. *Statistical procedures for agricultural research*. John Wiley & Sons, 1984.

Gregory M Hunter and Kenneth Steiglitz. Operations on images using quad trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):145–153, 1979.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

SL Lauritzen. Graphical models. volume 17 clarendon press, 1996.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Ian Mackay, Hans-Peter Piepho, and Antonio Augusto Franco Garcia. Statistical methods for plant breeding. In *Handbook of Statistical Genomics: Two Volume Set*, pages 501–520. Wiley Online Library, 2019.

Matin Qaim. Role of new plant breeding technologies for food security and sustainable agricultural development. *Applied Economic Perspectives and Policy*, 42(2):129–150, 2020.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

Nancy R Zhang and David O Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

# Extrapolation spatiale du risque de présence de Xylella fastidiosa basée sur XGBoost

Camille Portes [1] & Dino Ienco [2] & Edith Gabriel [3]

[1] *INRAE, BioSP, 84914 Avignon, France, camille.portes@inrae.fr*
[2] *UMR TETIS, 34000 Montpellier, France, dino.ienco@inrae.fr*
[3] *INRAE, BioSP, 84914 Avignon, France, edith.gabriel@inrae.fr*

**Résumé.** Nous proposons une méthodologie complète pour évaluer et cartographier le risque de présence de la bactérie *Xylella fastidiosa* en intégrant les composantes spatiales dans le processus de modélisation. Notre approche est basée sur l'apprentissage automatique, tenant compte des particularités des données : hétérogénéité spatiale et déséquilibre. Une pré-sélection de facteurs est réalisée à l'aide d'une approche ensembliste couplée à une validation croisée spatiale. Nous proposons ensuite une adaptation du modèle XGBoost dans laquelle les composantes spatiales sont intégrées au modèle, notamment en considérant des facteurs spatialement pondérés et en basant la sélection de modèle sur une validation croisée par blocs environnementaux. Cette approche nous permet d'obtenir un modèle robuste, généralisable à une zone géographique éloignée de celle utilisée pour l'entraînement du modèle et donc adaptée à l'extrapolation.

**Mots-clés.** Prédiction, Validation croisée spatiale, XGBoost

**Abstract.** We propose a methodology to assess and map the risk of presence of the bacterium *Xylella fastidiosa* by integrating spatial components into the modeling process. Our approach is based on machine learning, taking into account the data's specificities: spatial heterogeneity and imbalance. A pre-selection of factors is performed using an ensemble approach coupled with spatial cross-validation. We then propose an adapted XGBoost model where spatial components are integrated, both by considering spatially weighted factors and by basing model selection on block environmental cross-validation. This approach enables us to obtain a robust model that is generalizable to a geographic area distant from the one used for model training and thus suitable for extrapolation.

**Keywords.** Prediction, Spatial Cross Validation, XGBoost

## 1 Introduction

Connaître la santé des plantes dans une région permet de confirmer l'absence de la plupart des organismes nuisibles réglementés. La surveillance officielle repose sur des inspections visuelles, des prélèvements et des analyses d'échantillons. En cas de détection, les autorités sanitaires coordonnent des mesures de lutte collective pour empêcher l'établissement de l'organisme

nuisible dans une zone, préservant ainsi le reste du territoire. Nous nous intéressons ici à *Xylella fastidiosa*, bactérie vectée par des insectes (cicadelles) et potentiellement présente dans pas moins de 260 espèces de plantes. Apparue en Italie en 2013 et découverte plus tard en Espagne et au Portugal, la bactérie est désormais observée en Corse, en Provence-Alpes-Côte d'Azur et tout récemment en Occitanie. Dans ce contexte nos objectifs sont d'identifier les facteurs de risque de présence de la bactérie et de proposer un modèle d'extrapolation spatiale pour cartographier ce risque et orienter la surveillance.

Nous développons une méthodologie basée sur l'apprentissage automatique qui prend en compte les spécificités des données à chaque étape de la modélisation : hétérogénéité spatiale et déséquilibre (le nombre d'échantillons négatifs est largement supérieur aux échantillons positifs) et qui permet de faire des prédictions spatiales au-delà des emplacements des échantillons d'entraînement. Nous illustrerons l'approche en utilisant les données de Corse et de PACA comme ensemble d'entrainement et les données d'Occitanie comme ensemble de test.

## 2 Sélection de facteurs

Nous disposons d'un ensemble de 107 facteurs qui influencent le développement des plantes et donc la propagation des organismes nuisibles. Il s'agit de variables bioclimatiques (calculées à partir des variables climatiques de Siclima), de composition du sol (issues de la base Agroenvgeo), de type de sol (issues de Corine Land Cover), d'orientation du terrain et d'altitude. La base de données est décrite dans Portes *et al.* (2024).

Après élimination des facteurs trop fortement corrélés, nous sélectionnons les facteurs via une approche ensembliste couplée à une validation croisée spatiale (voir Section 3). Afin d'éviter d'éventuels biais d'induction, les modèles considérés sont de différent type : basé sur du boosting (e.g. adaboost, glmboost, xgboost...), sur des arbres (comme le random forest classique ou conditionnel), sur de l'analyse discriminante, sur les support vector machine, sur des modèles de régression logistique, ... Chacun des 29 modèles, adapté à notre contexte de classification binaire (plante testée positive versus négative), sélectionne 20 facteurs parmi nos 107 facteurs + 1 variable aléatoire. Cette sélection est faite en utilisant l'importance basée sur l'erreur "Out-of-Bag". Ainsi, nous avons considéré et comparé les résultats issus de différents critères :

1. Gain d'information

2. Importance (réduite) lorsqu'elle est calculable pour le modèle considéré ou poids uniforme lorsqu'elle ne l'est pas (dans ce cas le modèle sélectionne 20 facteurs mais leurs rangs sont inconnus),

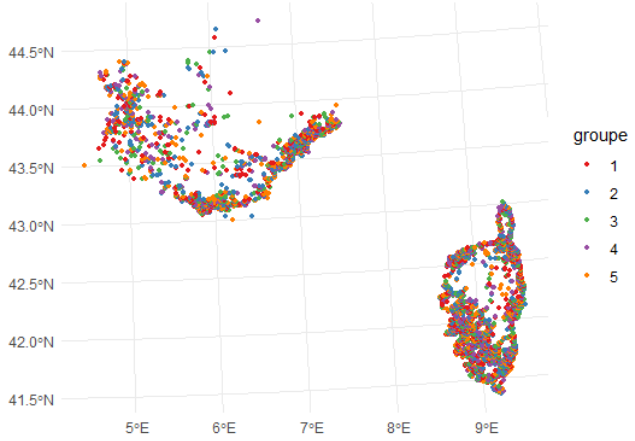
3. Importance (réduite) lorsqu'elle est calculable pour le modèle considéré.

Nous avons écarté le cas 1. car il s'est révélé être assez sensible à l'échelle du support d'observation de certains facteurs, en particulier pour la composition en éléments chimiques du sol. Les cas 2. et 3. mènent sensiblement aux mêmes résultats, avec 83 facteurs retenus (importance positive).

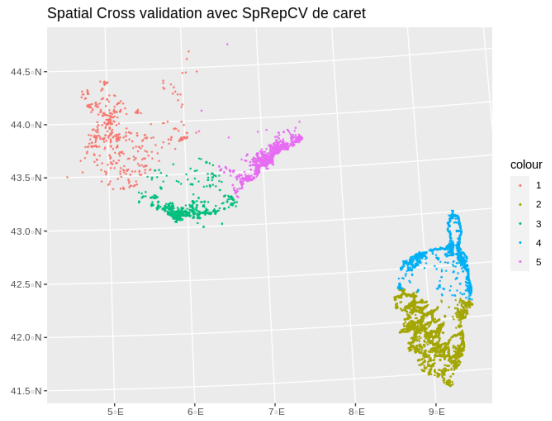
# 3 Validation croisée

Meyer et Pebesma (2021) mettent en évidence que, dans de nombreux contextes spatiaux, l'utilisation de la validation croisée aléatoire (Figure 1a) donne des résultats trop optimistes près des échantillons d'entrainement et de mauvais résultats en extrapolation et qu'il est préférable d'utiliser la validation croisée spatiale ou par blocs environnementaux (Valavi *et al.*, 2018).

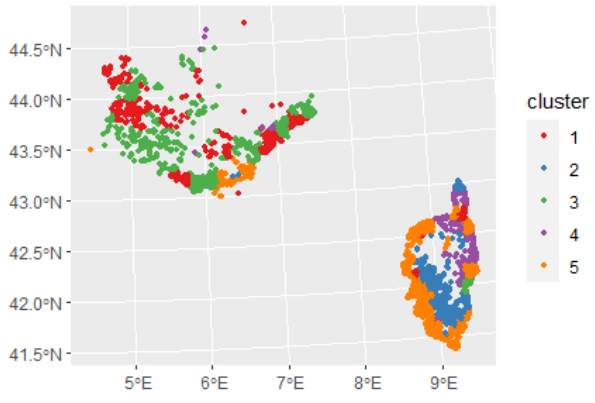
Dans un contexte spatial, la validation croisée valide le modèle sur un ensemble spatialement similaire (Figure 1b) à l'ensemble d'entraînement. Deux observations qui sont proches spatialement sont également proches en valeur pour les variables autocorrélées spatialement. Si le domaine de prédiction est éloigné, le modèle se comportera mal car il sera trop ajusté aux données d'entraînement et donc pas généralisé. La validation croisée spatiale contribue à résoudre ce problème mais peut ne pas être suffisante ou tout à fait adaptée.




(a) Validation croisée aléatoire




(b) Validation croisée spatiale

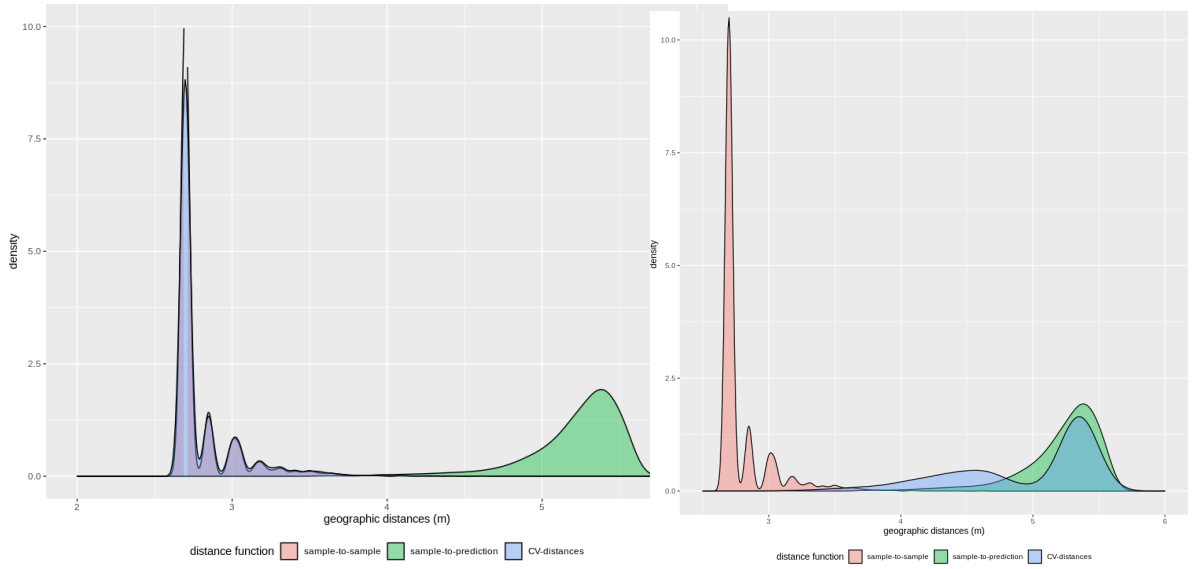



(c) Validation croisée par blocs environnementaux

Figure 1: Repésentation des ensembles d'entrainement et de validation selon le type de validation croisée.

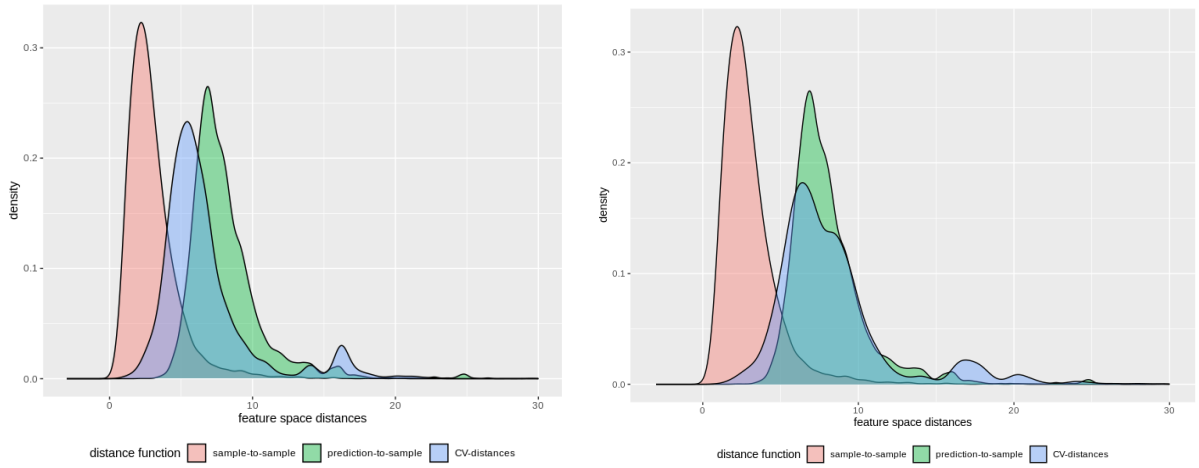
Pour gagner en robustesse sur la prédiction dans des zones éloignées, nous utilisons la validation croisée par blocs environnementaux (Figure 1c). Cette dernière utilise des méthodes de regroupement (k-means) pour spécifier des ensembles de facteurs similaires en fonction des variables d'entrée et du nombre de clusters choisi dans l'espace des facteurs.

La validation croisée par blocs environnementaux est adaptée à notre contexte. En effet, nous avons calculé les distances moyennes entre les points de l'ensemble d'entrainement et les distances entre les points de l'ensemble d'entrainement et de l'ensemble de test. Ces distances sont calculées dans un premier temps avec la position des observations, c'est donc une distance spatiale (Figure 2). Dans un deuxième temps, la distance est calculée à partir des facteurs (indice de dissimilarité) (Figure 3), qui est corrélé à la distance spatiale. La validation croisée spatiale, comme la validation croisée par blocs environnementaux, permet de valider le modèle sur un ensemble plus éloigné de l'ensemble d'entraînement (Figure 2b) comparé à la validation croisée aléatoire (Figure 2a). Cette distance spatiale se traduit par une dissimilarité de l'environnement, c'est-à-dire des facteurs. La validation croisée spatiale (Figure 3a) et la validation croisée par blocs environnementaux (Figure 3b) permettent d'avoir des écarts en terme de similarité du même ordre de grandeur entre les ensembles d'entraînement et de validation et les ensembles d'entraînement et de test, ce qui nous assure de la fiabilité du modèle, choisi sur l'ensemble de validation et sur l'ensemble de test.




(a) Validation croisée aléatoire (b) Validation croisée spatiale

Figure 2: Distances entre les ensembles d'entrainement, de validation et de test




(a) Validation croisée spatiale (b) Validation croisée par blocs environementaux

Figure 3: Similarités des environnements entre les ensembles d'entrainement, de validation et de test

# 4 Modélisation

L'autocorrélation spatiale et l'hétérogénéité spatiale sont deux effets spatiaux bien connus dans l'analyse spatiale et la modélisation. L'autocorrélation spatiale fait référence au processus qui crée des amas de valeurs. Dans notre cas, les amas locaux de cas positifs sont à la fois liés au mode de vection de la bactérie et au processus de surveillance (renforcé dans un périmètre étroit autour d'un cas positif). Nous nous affranchissons d'une grande partie de ce biais spatial en agrégeant les données sur une grille de maille 500 $m \times 500$ $m$ comme expliqué dans Martinetti et Soubeyrand (2019). L'hétérogénéité spatiale fait référence au fait que le processus qui génère les données peut différer d'un endroit à un autre. Elle est fréquemment modélisée par des modèles de coefficients variant spatialement (c'est par exemple le cas de la Régression Pondérée Géographiquement).

La littérature de cette dernière décennie montre un progrès visible dans le développement de la modélisation spatiale d'apprentissage automatique. Les modèles rencontrés se déclinent dans les combinaisons "ML classique ou spatial + variables non spatiales et/ou spatiales + validation croisée aléatoire ou spatiale". Le progrès et les innovations sont principalement liés à la formulation des méthodes d'apprentissage automatique pour incorporer des composantes spatiales dans les algorithmes.

Nous utilisons le modèle XGBoost (eXtreme Gradient Boosing). Cette méthode de boosting de gradient assemble séquentiellement des arbres de décision pour minimiser l'erreur du modèle en utilisant un algorithme d'optimisation de descente de gradient (Chen et Guestrin, 2016). La littérature montre qu'un XGBoost correctement ajusté surpasse généralement les méthodes alternatives telles que la forêt aléatoire ou le réseau de neurones profonds pour les problèmes supervisés.

Nous intégrons les composantes spatiales dans ce modèle d'une part en considérant, lorsque cela est pertinent, des facteurs spatialement pondérés et d'autre part en basant la sélection de modèle sur une validation croisée par blocs environnmentaux. Notre démarche est résumée dans la figure 4.

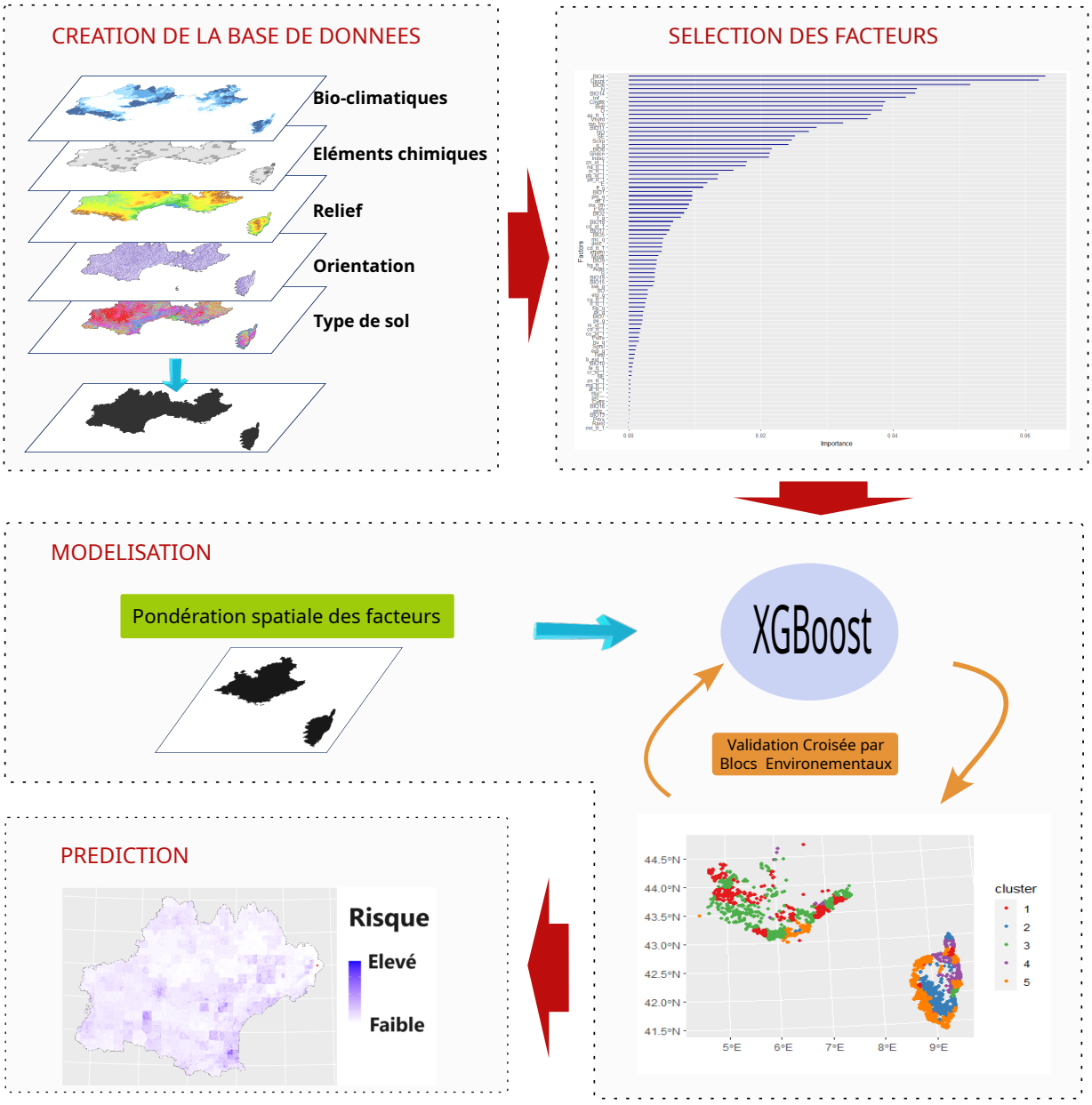



Figure 4: Méthodologie d'extrapolation spatiale basée sur XGBoost.

Ainsi, pour chacun des 83 facteurs sélectionnés nous avons testé l'existence d'une structure spatiale : 8 d'entre eux n'en ont pas. Pour les 75 autres, nous avons réalisé une analyse variographique et estimé la portée (distance au-delà de laquelle on considère qu'il n'y a plus de corrélation). Nous avons ensuite calculé la matrice de poids associée à chacun de ces facteurs à partir d'un noyau de type "triangle", avec une fenêtre égale à la portée, et telle

que seuls les 20 premiers voisins soient pris en compte.

Notre ensemble de facteurs $X$ est donc divisé en deux $X = (X_a, X_s)$ où $X_a$ désigne l'ensemble des facteurs aspatiaux et $X_s$ l'ensemble des facteurs spatiaux. L'algorithme a ensuite été appliqué à l'ensemble $(X_a, W_1X_{1,s}, \ldots, W_{75}X_{75,s})$ que l'on notera $WX$.

Pour comparer les modèles, nous utilisons des mesures qui prennent en compte le déséquilibre du ratio positifs/négatifs, telles que la balanced accuracy (qui calcule le pourcentage de positifs et négatifs bien prédits parmi tous les positifs et tous les négatifs), le F1-score (qui calcule la moyenne entre le pourcentage de positifs bien prédits parmi les positifs et le pourcentage de positifs parmi ceux prédits comme positifs) ou l'AUC.

Nous illustrerons la comparaison des résultats obtenus à partir des modèles XGBoost($X$), XGBoost($WX$) et XGBoost($X, WX$) et pour les validations croisées aléatoire, spatiale et par blocs environnementaux.

# Bibliographie


Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. pages 785–794.

Martinetti, D. and Soubeyrand, S. (2019). Identifying lookouts for epidemio-surveillance: application to the emergence of xylella fastidiosa in france. Phytopathology, 109(2):265–276.

Meyer, H. and Pebesma, E. (2021). Predicting into unknown space? estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution, 12(9):1620–163.

Portes, C., Ienco, D. and Gabriel, E. (2024). Environmental and Bio-climatic Data for Epidemiological Analysis over French Mediterranean areas. (Soumis).

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2018). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. bioRxiv, page 357798.

# Classification

# Projections Aléatoires Entrée, Sortie : Accélération de l'Apprentissage et de l'Inférence dans la Prédiction Structurée avec Noyaux

Tamim El Ahmad [1] & Luc Brogat-Motte [2] & Pierre Laforgue [3] & Florence d'Alché-Buc [4]

[1] *LTCI, Télécom Paris, IP Paris, France, tamim.elahmad@telecom-paris.fr*
[2] *L2S, CentraleSupélec, France, luc.brogat.motte@l2s.centralesupelec.fr*
[3] *Department of Computer Science, University of Milan, Italy, pierre.laforgue@unimi.it*
[4] *LTCI, Télécom Paris, IP Paris, France, florence.dalche@telecom-paris.fr*

**Résumé.** Grâce à l'utilisation de l'astuce du noyau dans les espaces d'entrée et de sorties, les méthodes subrogées à noyaux offrent une solution polyvalente avec des fondements théoriques au problème de prédiction structurée. Bien que sur des ensembles de données de tailles modérées elles constituent l'état de l'art, comme dans la chimie-informatique par exemple, elles échouent à passer à l'échelle lorsque le nombre de données d'entraînement devient élevé. Nous proposons d'équiper ces méthodes subrogées à noyaux avec des approximations à l'aide de projections aléatoires, appliquées aux noyaux d'entrée et de sortie. Nous prouvons une borne d'excès de risque sur l'estimateur du problème structuré, atteignant une vitesse de convergence proche de l'estimateur optimal avec des projecteurs de petite dimension en fonction des vitesses de décroissance des valeurs propres des opérateurs de covariance entrée/sortie. D'un point de vue computationnel, nous montrons que les deux approximations ont des impacts distincts mais complémentaires : en entrée on accélère principalement l'apprentissage, tandis qu'en sortie c'est l'inférence qui est accélérée.

**Mots-clés.** Méthodes à noyaux, Projections aléatoires, Apprentissage machine à grande échelle, Prédiction Structurée, Apprentissage statistique.

**Abstract.** Leveraging the kernel trick in both the input and output spaces, surrogate kernel methods are a flexible and theoretically grounded solution to structured output prediction. If they provide state-of-the-art performance on complex data sets of moderate size (e.g., in chemoinformatics), these approaches however fail to scale. We propose to equip surrogate kernel methods with sketching-based approximations, applied to both the input and output feature maps. We prove excess risk bounds on the original structured prediction problem, showing how to attain close-to-optimal rates with a reduced sketch size that depends on the eigendecay of the input/output covariance operators. From a computational perspective, we show that the two approximations have distinct but complementary impacts: sketching the input kernel mostly reduces training time, while sketching the output kernel decreases the inference time.

**Keywords.** Kernel methods, Sketching, Large-scale Machine Learning, Structured Prediction, Statistical learning.

# 1 Introduction

Ubiquitous in real-world applications, structured objects have attracted a great deal of attention in machine learning (Bakir et al., 2007; Gärtner, 2008; Nowozin and Lampert, 2011; Deshwal et al., 2019). Depending on their role, i.e., either as input or output variables, they raise distinct challenges. Classification and regression from structured *inputs* generally rely on a continuous representation learned by a deep neural network (Defferrard et al., 2016), or implicitly defined through a dedicated kernel (Collins and Duffy, 2001; Borgwardt et al., 2020). In contrast, structured *output* prediction calls for a more involved approach, since the discrete nature of the outputs impacts the definition of the loss function (Nowak et al., 2019; Ciliberto et al., 2020; Cabannes et al., 2021), and therefore the learning problem itself.

To handle this problem, several methods have been developed to relax the combinatorial problems that appear both at training and inference. Energy-based approaches convert structured prediction into learning a scalar score function (Tsochantaridis et al., 2005; LeCun et al., 2007; Belanger and McCallum, 2016; Deshwal et al., 2019). End-to-end learning typically exploits a differentiable model, together with a differentiable loss, to run gradient descent (Long et al., 2015; Niculae et al., 2018; Berthet et al., 2020). Surrogate methods (Ciliberto et al., 2020) solve a regression problem in a Hilbert space where outputs have been implicitly embedded, shortcutting the inference during learning.

Rare are the methods that enjoy both scalability at learning/inference steps and statistical guarantees (Osokin et al., 2017; Cabannes et al., 2021). In this work, we focus on surrogate approaches and their implementation as kernel methods, i.e., the input output kernel regression framework (Cortes et al., 2005; Brouard et al., 2016b). Recent works Ciliberto et al. (2016, 2020) have shown that they enjoy consistency, their excess risk being governed by that of the surrogate regression. Moreover, they are well appropriate to make prediction from one structured modality to another, since kernels can be leveraged in both the input and output spaces. Overall, they offer a general, theoretically grounded, and simple-to-implement solution to structured prediction, providing state-of-the-art results in applications such as molecule identification (Schymanski et al., 2017).

However, contrary to deep neural networks, they do not scale neither in memory nor in time without further approximation. The aim of this paper is to equip these methods with kernel approximations to obtain a drastic complexity reduction while maintaining their statistical properties. Several works have highlighted the power of kernel approximations, from Random Fourier Features (Rahimi and Recht, 2007; Brault et al., 2016; Rudi and Rosasco, 2017; Li et al., 2021), to general low-rank approaches (Bach, 2013; Meanti et al., 2020).

In this work we focus on sketching (Mahoney et al., 2011; Woodruff, 2014), a general dimension reduction method based on linear random projections. Applied to kernel approximation, sketching has been widely studied through Nyström's sub-sampling approximation (Williams and Seeger, 2001; Alaoui and Mahoney, 2015; Rudi et al., 2015), and further explored using Gaussian or Randomized Orthogonal Systems (Yang et al., 2017; Lacotte and Pilanci, 2020). Interpreted as a way to provide data-dependent random features (Williams and Seeger, 2001; Yang et al., 2012; Kpotufe and Sriperumbudur, 2020), this approach has allowed to scale up kernel PCA (Sterge and Sriperumbudur, 2022), kernel mean embedding

(Chatalic et al., 2022a,b) or independence tests (Kalinke and Szabó, 2023) while enjoying statistical guarantees. However, sketching has been limited so far to scalar kernel machines. No current approach covers both sides of the coin, i.e., applying approximations to both the input and output kernels. Motivated by surrogate structured prediction, we close this gap and make the following contributions:

- We apply sketching to the vector-valued kernel regression problem solved in structured prediction, both on inputs and outputs, which accelerates respectively learning and inference.
- We derive excess risk bounds controlled by the properties of the sketched projection operators.
- We prove that sub-Gaussian sketches provide close-to-optimal rates with small sketch sizes.
- We empirically show that our algorithms maintain good accuracy on moderate size datasets, while enabling kernel surrogate methods on large datasets where the standard approach is simply intractable.

**Notations.** We introduce now generic notations for the input (output) space and kernel. If $\mathcal{Z}$ denotes a generic Polish space, $\mathrm{k}_{\mathcal{Z}}$ is a positive definite kernel over $\mathcal{Z}$ and $\psi_{\mathcal{Z}}(z) \coloneqq \mathrm{k}_{\mathcal{Z}}(\cdot, z)$ is the canonical feature map of $\mathrm{k}_{\mathcal{Z}}$. $\mathcal{H}_{\mathcal{Z}}$ denotes the Reproducing Kernel Hilbert Space (RKHS) associated to $\mathrm{k}_{\mathcal{Z}}$. $\mathrm{S}_{\mathrm{Z}} : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \ldots, f(z_n))^\top$ is the sampling operator over $\mathcal{H}_{\mathcal{Z}}$ (Smale and Zhou, 2007). For any operator $A$, we denote $A^\#$ its adjoint. The adjoint of $\mathrm{S}_{\mathrm{Z}}$ is defined as ${\mathrm{S}_{\mathrm{Z}}}^\# : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \, \psi_{\mathcal{Z}}(z_i)$. If $z$ is a r.v. distributed according to $\rho_{\mathcal{Z}}$, its covariance operator over $\mathcal{H}_Z$ is $\mathrm{C}_{\mathcal{Z}} = \mathbb{E}_z[\psi_{\mathcal{Z}}(z) \otimes \psi_{\mathcal{Z}}(z)]$, and its empirical counterpart $\widehat{\mathrm{C}}_{\mathrm{Z}} = (1/n) \sum_{i=1}^n \psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i) = {\mathrm{S}_{\mathrm{Z}}}^\# \mathrm{S}_{\mathrm{Z}}$, where $\{(z_i)_{i=1}^n\}$ is i.i.d. drawn from $\rho_{\mathcal{Z}}$. The Moore-Penrose inverse of $M$ is denoted $M^\dagger$.

## 2 Background

We now recall the structured prediction setting based on a kernel-induced loss, and a state-of-the-art surrogate approach to solve it. We also provide reminders about sketching as a way to scale-up kernel methods.

**Structured prediction with surrogate kernel methods.** Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ a structured output space. In general, $\mathcal{Y}$ is finite and extremely large. Let a positive definitew kernel $\mathrm{k}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, that measures how close two objects from $\mathcal{Y}$ are. We consider the loss function induced by $\mathrm{k}_{\mathcal{Y}}$, defined as $\ell : (y, y') \to \| \psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(y') \|^2_{\mathcal{H}_{\mathcal{Y}}}$. Note that it can be computed using the kernel trick. Given an unknown joint probability distribution $\rho$ defined on $\mathcal{X} \times \mathcal{Y}$, the goal of structured prediction is to approximate

$$f^* = \operatorname*{arg\,min}_{f : \mathcal{X} \to \mathcal{Y}} \mathcal{R}(f) \,, \tag{1}$$

where $\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\rho}\left[\| \psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(f(x))\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]$, using only an i.i.d. sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from $\rho$. Estimating directly $f^*$ is not tractable, such that many works (Cortes et al., 2005; Geurts et al., 2006; Brouard et al., 2011; Ciliberto et al., 2016) have proposed instead the following two-step approach:

**1. Surrogate Regression:** Find an estimator $\hat{h}$ of the surrogate target $h^*\colon x \mapsto \mathbb{E}_y[\psi_{\mathcal{Y}}(y)|x]$ such that

$$h^* = \arg\min_h \ \mathbb{E}_{(x,y)}\left[\|h(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right].$$

**2. Pre-image:** Define $\hat{f}$ by decoding $\hat{h}$, i.e.,

$$\hat{f}(x) = d(\hat{h}(x)) \coloneqq \arg\min_{y\in\mathcal{Y}} \left\|\hat{h}(x) - \psi_{\mathcal{Y}}(y)\right\|^2_{\mathcal{H}_{\mathcal{Y}}}.$$

The surrogate regression in Step 1 is much easier to handle than the initial structured prediction problem: it avoids learning $f$ through the composition with the implicit feature map $\psi_{\mathcal{Y}}$, and relegates the difficulty of handling structured objects to Step 2, i.e. at inference. In addition, vector-valued regression into infinite-dimensional spaces is a well-studied problem, that can be solved by using the kernel trick in the output space. This two-step approach belongs to the general framework of SELF (Ciliberto et al., 2016) and ILE (Ciliberto et al., 2020) and enjoys valuable theoretical guarantees. It is Fisher consistent, i.e., $h^*$ yields $f^*$ after decoding, and the excess risk of $\hat{f}$ is controlled by that of $\hat{h}$.

**Input Output ridge Kernel Regression.** A common choice to tackle in practice the surrogate regression problem consists in solving a *kernel ridge regression problem*, leveraging kernels in both input and output spaces. The hypothesis space is chosen as a vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) (Senkene and Tempel'man, 1973; Micchelli and Pontil, 2005; Carmeli et al., 2006, 2010). In the same way that RKHS are based on positive symmetric definite kernels, vv-RKHS are based on Operator-Valued Kernels (OVK). In our setting, we define an OVK $\mathcal{K}$, as a mapping $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_{\mathcal{Y}})$, where $\mathcal{L}(\mathcal{H}_{\mathcal{Y}})$ is the set of bounded linear operators on $\mathcal{H}_{\mathcal{Y}}$, and that satisfies the properties recalled in Appendix B. An OVK $\mathcal{K}$ is uniquely associated with a vv-RKHS $\mathcal{H}$, i.e. a Hilbert space of functions from $\mathcal{X}$ to $\mathcal{H}_{\mathcal{Y}}$ that enjoys the reproducing kernel property (see Appendix B).

In what follows, we opt for the identity decomposable OVK $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_{\mathcal{Y}})$, defined as: $\mathcal{K}(x, x') = \mathrm{k}_{\mathcal{X}}(x, x')\, I_{\mathcal{H}_{\mathcal{Y}}}$, where $\mathrm{k}_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a p.d. scalar-valued kernel on $\mathcal{X}$. In *Input Output Kernel Ridge Regression* (IOKR for short, Brouard et al. 2011, 2016b; Ciliberto et al. 2020, also introduced as Kernel Dependency Estimation by Weston et al. (2003)), the estimator of the surrogate regression is obtained by solving the following Ridge regression problem within $\mathcal{H}$, given a regularisation penalty $\lambda > 0$,

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \ \frac{1}{n}\sum_{i=1}^{n} \|\psi_{\mathcal{Y}}(y_i) - h(x_i)\|^2_{\mathcal{H}_{\mathcal{Y}}} + \lambda\|h\|^2_{\mathcal{H}}. \tag{2}$$

Interestingly, the unique solution to the above problem can be expressed in different ways. From one hand, we can derive from the representer theorem in vv-RKHSs (Micchelli and

Pontil, 2005) the following expression:

$$\hat{h}(x) = \sum_{i=1}^{n} \hat{\alpha}_i(x)\, \psi_{\mathcal{Y}}(y_i)\,, \quad \text{with} \quad \hat{\alpha}(x) = (\mathrm{K_X} + n\lambda)^{-1}\, \mathrm{k_X^x} := \widehat{\Omega}\, \mathrm{k_X^x}\,, \tag{3}$$

where $\mathrm{K_X} = (\mathrm{k}_{\mathcal{X}}(x_i, x_j))_{i,j=1}^{n}$ and $\mathrm{k_X^x} = \big(\mathrm{k}_{\mathcal{X}}(x, x_1), \ldots, \mathrm{k}_{\mathcal{X}}(x, x_n)\big)$. On the other hand, using an operator view one obtains

$$\hat{h}(x) = \widehat{H}\, \psi_{\mathcal{X}}(x)\,, \quad \text{where} \quad \widehat{H} = \mathrm{S_Y}^{\#}\, \mathrm{S_X}\, \big(\widehat{\mathrm{C}}_{\mathrm{X}} + \lambda I\big)^{-1}\,. \tag{4}$$

The latter expression can be seen as a re-writing of the first (Ciliberto et al., 2016), echoing the KDE equations with finite-dimensional feature maps (Cortes et al., 2005). It can also be related to the conditional kernel empirical mean embedding (Grünewälder et al., 2012).

The final estimator $\hat{f}$ is computed using the expression in (3), in order to benefit from the kernel trick:

$$\hat{f}(x) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \; \mathrm{k}_{\mathcal{Y}}(y, y) - 2\mathrm{k_X^x}^{T} \widehat{\Omega} \mathrm{k_Y^y}\,, \tag{5}$$

where $\mathrm{k_Y^y} = (\mathrm{k}_{\mathcal{Y}}(y, y_1), \ldots, \mathrm{k}_{\mathcal{Y}}(y, y_n))^{\top}$. The training phase thus involves the inversion of a $n \times n$ matrix, whose cost without any approximation is $\mathcal{O}(n^3)$. Besides, it implies storing $n^2$ values in memory, which induces a heavy space complexity as well. In practice, decoding is performed by searching in a candidate set $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size $n_c$. Hence, performing predictions on a test set $X_{\mathrm{te}}$ of size $n_{\mathrm{te}}$ mainly implies computing

$$\underbrace{\mathrm{K_X}^{\mathrm{te,tr}}}_{n_{\mathrm{te}} \times n}\, \underbrace{\widehat{\Omega}}_{n \times n}\, \underbrace{\mathrm{K_Y}^{\mathrm{tr},c}}_{n \times n_c}, \tag{6}$$

where $\mathrm{K_X}^{\mathrm{te,tr}} = (\mathrm{k}_{\mathcal{X}}(x_i^{\mathrm{te}}, x_j))_{1 \le i \le n_{\mathrm{te}}, 1 \le j \le n} \in \mathbb{R}^{n_{\mathrm{te}} \times n}$, and $\mathrm{K_Y}^{\mathrm{tr},c} = \big(\mathrm{k}_{\mathcal{Y}}(y_i, y_j^c)\big)_{1 \le i \le n, 1 \le j \le n_c} \in \mathbb{R}^{n \times n_c}$. The complexity of the decoding part is $\mathcal{O}\left(n_{\mathrm{te}} n n_c\right)$, considering $n_{\mathrm{te}} < n \le n_c$. IOKR thus suffers from both heavy time and space computational costs. To cope with this limitation, we develop a general sketching approach that applies to both input and output feature spaces, accelerating both training and decoding.

**Sketching for kernel methods.** Applied to kernel methods to reduce their dependency in $n$, sketching can be seen as linear projections induced by a random matrix $R$ (the sketching matrix) drawn from a probability distribution over $\mathbb{R}^{m \times n}$, where $m \ll n$. Classic examples include Nyström's approximation, where each row of $R$ is randomly drawn from the rows of the identity matrix $I_n$, and Gaussian sketches, where all entries of $R$ are i.i.d. Gaussian random variables. Nyström's approximation acts as a random training data sub-sampler, but it can be interpreted in many ways. In Drineas et al. (2005); Bach (2013), it is shown to generate a low-rank approximation of the Gram matrix, while in Williams and Seeger (2001); Yang et al. (2012), it is seen as a way to construct data-dependent finite-dimensional random features. In Rudi et al. (2015), instead, it is presented as a projection onto a small subspace of the RKHS. For other sketching schemes such as Gaussian or Randomized Orthogonal Systems, most of the works adopt an optimization viewpoint, where a variable substitution is operated

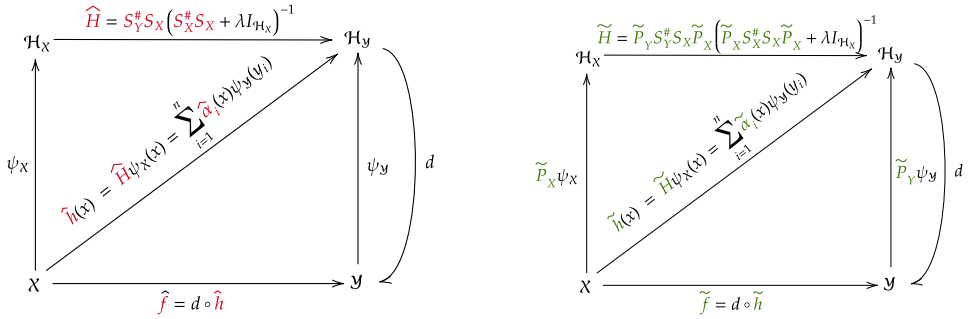


Figure 1: IOKR (left) and SISOKR (right) in the KDE setting. Note that SISOKR consists in IOKR when kernels $\mathrm{k}_{\mathcal{Z}}$ are replaced with their projected versions $\tilde{k}_{\mathcal{Z}}(\cdot,\cdot) = \langle \psi_{\mathcal{Z}}(\cdot), \widetilde{\mathrm{P}}_{\mathrm{Z}}\, \psi_{\mathcal{Z}}(\cdot)\rangle_{\mathcal{H}_{\mathcal{Z}}}$. However, this new output kernel changes the pre-image problem, and consequently the estimator $\tilde{f}$. In the paper, we modify $\widetilde{H}$ (and not the kernels) in order to use the comparison inequality from Ciliberto et al. (2020), see the proof of Corollary 1.

after the application of a Representer theorem (Yang et al., 2017; Lacotte and Pilanci, 2020). An interesting view provided in Kpotufe and Sriperumbudur (2020) explores the construction of random features based on Gaussian sketching. All these works are however limited to sketching the *input* kernel, in scalar regression problems. In this work: (1) we generalize input sketching to vector-valued problems, (2) we sketch the outputs, which is critical to scale-up surrogate methods with kernelized outputs.

# 3 Sketched Input Sketched Output Kernel Regression

The goal of this section is to construct a low-rank estimator of $\hat{h}$ by using sketching on both the input and output kernels. Note that sketching the feature maps is not desirable here: if we replace the output features $\psi_{\mathcal{Y}}(y_i) \in \mathcal{H}_{\mathcal{Y}}$ with some sketch-dependent approximations $\tilde{\psi}_{\mathcal{Y}}(y_i) \in \mathbb{R}^m$ we become unable to compare the resulting $\tilde{h}$ to the target $h^*$. Indeed, $\tilde{h}$ is an approximation of $x \mapsto \mathbb{E}_y[\tilde{\psi}_{\mathcal{Y}}(y)|x]$, which is a biased version of $h^*$ due to the sketch realization. Instead, as we show below, seeing sketching as orthogonal projections provides a natural way to solve our problem. Ultimately, this gives rise to an estimator $\tilde{f}$ for structured prediction which is versatile, easy-to-implement, theoretically-based and scalable to large data sets.

**Low-rank estimator.** Given two orthogonal projection operators $\widetilde{\mathrm{P}}_{\mathrm{X}}$ and $\widetilde{\mathrm{P}}_{\mathrm{Y}}$, we start from (4) and replace the sampling operators on both sides, $\mathrm{S}_{\mathrm{X}}$ and $\mathrm{S}_{\mathrm{Y}}$, by their projected counterparts, $\mathrm{S}_{\mathrm{X}}\,\widetilde{\mathrm{P}}_{\mathrm{X}}$ and $\mathrm{S}_{\mathrm{Y}}\,\widetilde{\mathrm{P}}_{\mathrm{Y}}$, so as to encode dimension reduction. The proposed low-rank estimator is expressed as follows:

$$\tilde{h}(x) = \widetilde{\mathrm{P}}_{\mathrm{Y}}\, \mathrm{S}_{\mathrm{Y}}{}^{\#}\, \mathrm{S}_{\mathrm{X}}\, \widetilde{\mathrm{P}}_{\mathrm{X}} \left( \widetilde{\mathrm{P}}_{\mathrm{X}}\, \widehat{\mathrm{C}}_{\mathrm{X}}\, \widetilde{\mathrm{P}}_{\mathrm{X}} + \lambda I_{\mathcal{H}_{\mathcal{X}}} \right)^{-1} \psi_{\mathcal{X}}(x)\,.$$

We now show how to design the projection operators using sketching and then derive the novel expression of the low-rank estimator in terms of a weighted combination of the training

outputs: $\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i \, \psi_{\mathcal{Y}}(y_i)$, yielding a reduced computational cost. IOKR and SISOKR approaches are illustrated on Figure 1.

**Sketching.** In this work, we chose to leverage sketching to obtain random projectors within the input and output feature spaces. Indeed, sketching consists of approximating a feature map $\psi_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{H}_{\mathcal{Z}}$ by projecting it thanks to a random projection operator $\widetilde{\mathrm{P}}_{\mathrm{Z}}$ defined as follows. Given a random matrix $\mathrm{R}_{\mathcal{Z}} \in \mathbb{R}^{\mathrm{m}_{\mathcal{Z}} \times n}$, $n$ data $(z_i)_{i=1}^{n} \in \mathcal{Z}$ and $\mathrm{m}_{\mathcal{Z}} \ll n$, the linear subspace defining $\widetilde{\mathrm{P}}_{\mathrm{Z}}$ is constructed as the linear subspace generated by the span of the following $\mathrm{m}_{\mathcal{Z}}$ random vectors

$$\sum_{j=1}^{n} (\mathrm{R}_{\mathcal{Z}})_{ij} \, \psi_{\mathcal{Z}}(z_j) \in \mathcal{H}_{\mathcal{Z}}, \quad i = 1, \ldots, \mathrm{m}_{\mathcal{Z}} \; .$$

One can show (Proposition 2 in Appendix C) that the corresponding orthogonal projector writes

$$\widetilde{\mathrm{P}}_{\mathrm{Z}} = (\mathrm{R}_{\mathcal{Z}} \, \mathrm{S}_{\mathrm{Z}})^{\#} \left( \mathrm{R}_{\mathcal{Z}} \, \mathrm{S}_{\mathrm{Z}} (\mathrm{R}_{\mathcal{Z}} \, \mathrm{S}_{\mathrm{Z}})^{\#} \right)^{\dagger} \mathrm{R}_{\mathcal{Z}} \, \mathrm{S}_{\mathrm{Z}} \; . \tag{7}$$

**Sketched Input Sketched Output Kernel Regression (SISOKR).** The SISOKR estimator is the low-rank estimator $\tilde{h}$, where both $\widetilde{\mathrm{P}}_{\mathrm{X}}$ and $\widetilde{\mathrm{P}}_{\mathrm{Y}}$ have been chosen as (7), for some random sketches $\mathrm{R}_{\mathcal{X}}$ and $\mathrm{R}_{\mathcal{Y}}$. It also admits the following expression based on a linear combination of the $\psi_{\mathcal{Y}}(y_i)$.

**Proposition 1** (Expression of SISOKR). *$\forall \, x \in \mathcal{X}$, $\tilde{h}\,(x) = \sum_{i=1}^{n} \tilde{\alpha}_i \,(x) \, \psi_{\mathcal{Y}}\,(y_i)$ where*

$$\tilde{\alpha}\,(x) = \mathrm{R}_{\mathcal{Y}}{}^{\top} \, \widetilde{\Omega} \, \mathrm{R}_{\mathcal{X}} \, \mathrm{k}_{\mathrm{X}}^{\mathrm{x}} \quad \textit{and} \quad \widetilde{\Omega} = \widetilde{\mathrm{K}}_{\mathrm{Y}}{}^{\dagger} \, \mathrm{R}_{\mathcal{Y}} \, \mathrm{K}_{\mathrm{Y}} \, \mathrm{K}_{\mathrm{X}} \, \mathrm{R}_{\mathcal{X}}{}^{\top} (\mathrm{R}_{\mathcal{X}} \, \mathrm{K}_{\mathrm{X}}{}^{2} \, \mathrm{R}_{\mathcal{X}}{}^{\top} + n\lambda \, \widetilde{\mathrm{K}}_{\mathrm{X}})^{\dagger} \, , \tag{8}$$

*with* $\widetilde{\mathrm{K}}_{\mathrm{X}} = \mathrm{R}_{\mathcal{X}} \, \mathrm{K}_{\mathrm{X}} \, \mathrm{R}_{\mathcal{X}}{}^{\top}$ *and* $\widetilde{\mathrm{K}}_{\mathrm{Y}} = \mathrm{R}_{\mathcal{Y}} \, \mathrm{K}_{\mathrm{Y}} \, \mathrm{R}_{\mathcal{Y}}{}^{\top}$.

Note that the matrix quantity that we recover above, $\mathrm{K}_{\mathrm{X}} \, \mathrm{R}_{\mathcal{X}}{}^{\top} \left( \mathrm{R}_{\mathcal{X}} \, \mathrm{K}_{\mathrm{X}}{}^{2} \, \mathrm{R}_{\mathcal{X}}{}^{\top} + n\lambda \, \widetilde{\mathrm{K}}_{\mathrm{X}} \right)^{\dagger} \mathrm{R}_{\mathcal{X}} \, \mathrm{k}_{\mathrm{X}}^{\mathrm{x}}$, is typical to sketched kernel Ridge regression (Rudi et al., 2015; Yang et al., 2017). It allows to reduce the size of the matrix to invert, which is now an $\mathrm{m}_{\mathcal{X}} \times \mathrm{m}_{\mathcal{X}}$ matrix. This is the main reason for the reduction of the learning step's complexity, and is due to the input sketching (still, we need to perform matrix multiplication $\mathrm{R}_{\mathcal{X}} \, \mathrm{K}_{\mathrm{X}}$, whose efficiency depends on the sketch used). Note that output sketching also requires additional operations, but the overall cost of computing $\tilde{\alpha}$ remains negligible compared to $\mathcal{O}(n^3)$. We obtain the corresponding structured prediction estimator $\tilde{f}$ by decoding $\tilde{h}$, i.e., by replacing $\widehat{\Omega}$ by $\widetilde{\Omega}$ in (5). In fact, the main quantity we have to compute for prediction is now

$$\underbrace{\mathrm{K}_{\mathrm{X}}{}^{\mathrm{te,tr}} \, \mathrm{R}_{\mathcal{X}}{}^{\top}}_{n_{\mathrm{te}} \times \mathrm{m}_{\mathcal{X}}} \; \underbrace{\widetilde{\Omega}}_{\mathrm{m}_{\mathcal{X}} \times \mathrm{m}_{\mathcal{Y}}} \; \underbrace{\mathrm{R}_{\mathcal{Y}} \, \mathrm{K}_{\mathrm{Y}}{}^{\mathrm{tr},c}}_{\mathrm{m}_{\mathcal{Y}} \times n_c} \; . \tag{9}$$

The time complexity of this operation is $\mathcal{O}(n_{\mathrm{te}} m_{\mathcal{Y}} n_c)$ if $n_{\mathrm{te}} \leq \mathrm{m}_{\mathcal{X}}, \mathrm{m}_{\mathcal{Y}} < n \leq n_c$, which is a significant complexity reduction (the dependence in $n$ vanishes), governed by the output sketch size $\mathrm{m}_{\mathcal{Y}}$, see Appendix I.

# 4 Theoretical Analysis

In this section, we present a statistical analysis of the proposed estimators $\tilde{h}$ and $\tilde{f}$. After introducing the assumptions on the learning task, we upper bound the excess-risk of the sketched kernel ridge estimator, highlighting the approximation errors due to sketching. We then provide bounds for these approximation error terms. Finally, we study under which setting the proposed estimators $\tilde{h}$ and $\tilde{f}$ obtain substantial computational gains, while still benefiting from a close-to-optimal learning rates. We consider the following set of common assumptions in the kernel literature (Bauer et al., 2007; Steinwart et al., 2009; Rudi et al., 2015; Pillaud-Vivien et al., 2018; Fischer and Steinwart, 2020; Ciliberto et al., 2020; Brogat-Motte et al., 2022).

**Assumption 1** (Attainability). *We assume that $h^* \in \mathcal{H}$, i.e., that there is a linear operator $H : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$, with $\|H\|_{\text{HS}} < +\infty$, s.t. $h^*(x) = H\,\psi_\mathcal{X}(x)$, $\forall\, x \in \mathcal{X}$.*

This is a standard assumption in the context of least-squares regression (Caponnetto and De Vito, 2007), making the target $h^*$ belong to the hypothesis space. Note that relaxing this assumption is possible, although it would add a bias term that still requires some knowledge about $h^*$ to be bounded. For instance, if $h^*$ is supposed to be square-integrable, one usually chooses a RKHS associated with a universal operator-valued kernel, which is dense in the space of the square-integrable functions (Carmeli et al., 2010, Section 4). We now describe a set of generic assumptions that have to be satisfied by both input and output kernels $\mathrm{k}_\mathcal{X}$ and $\mathrm{k}_\mathcal{Y}$.

**Assumption 2** (Bounded kernel). *There exists $\kappa_\mathcal{Z} > 0$ such that $\mathrm{k}_\mathcal{Z}(z, z) \leq {\kappa_\mathcal{Z}}^2$, $\forall\, z \in \mathcal{Z}$. We note $\kappa_\mathcal{X}, \kappa_\mathcal{Y} > 0$ for the input and output kernels $\mathrm{k}_\mathcal{X}$ and $\mathrm{k}_\mathcal{Y}$ respectively.*

**Assumption 3** (Capacity condition). *There exists $\gamma_\mathcal{Z} \in [0, 1]$ such that $\mathrm{Q}_\mathcal{Z} := \mathrm{Tr}({\mathrm{C}_\mathcal{Z}}^{\gamma_\mathcal{Z}}) < +\infty$.*

Note that Assumption 3 is always verified for $\gamma_\mathcal{Z} = 1$, as $\mathrm{Tr}(\mathrm{C}_\mathcal{Z}) = \mathbb{E}[\|\,\psi_\mathcal{Z}(z)\|^2_{\mathcal{H}_\mathcal{Z}}] < +\infty$ from Assumption 2, and that the smaller $\gamma_\mathcal{Z}$ the faster the eigendecay of $\mathrm{C}_\mathcal{Z}$, with $\gamma_\mathcal{Z} = 0$ when $\mathrm{C}_\mathcal{Z}$ is of finite rank. More generally, this assumption is for instance verified for a Sobolev kernel and a marginal distribution whose density is upper-bounded (Ciliberto et al., 2020, Assumption 2).

**Assumption 4** (Embedding property). *There exist $\mathrm{b}_\mathcal{Z} > 0$ and $\mu_\mathcal{Z} \in [0, 1]$ such that $\psi_\mathcal{Z}(z) \otimes \psi_\mathcal{Z}(z) \preceq \mathrm{b}_\mathcal{Z}\, {\mathrm{C}_\mathcal{Z}}^{1-\mu_\mathcal{Z}}$ almost surely.*

Note that Assumption 4 is always verified for $\mu_\mathcal{Z} = 1$, as $\psi_\mathcal{Z}(z) \otimes \psi_\mathcal{Z}(z) \preceq {\kappa_\mathcal{Z}}^2\, I_{\mathcal{H}_\mathcal{Z}}$ by Assumption 2, and that the smaller $\mu_\mathcal{Z}$, the stronger the assumption, with $\mu_\mathcal{Z} = 0$ when $\mathrm{C}_\mathcal{Z}$ is of finite. It allows to control the regularity of the functions in $\mathcal{H}_\mathcal{Z}$ with respect to the $L^\infty$-norm, as it implies $\|h\|_{L_\infty} \leq {\mathrm{b}_\mathcal{Z}}^{1/2}\, \|h\|^\mu_{\mathcal{H}_\mathcal{Z}}\, \mathbb{E}[h(z)^2]^{(1-\mu)/2}$ (Pillaud-Vivien et al., 2018). For instance, an absolutely continuous distribution whose density is lower-bounded almost everywhere and a Matérn kernel verifies Assumption 4 (Pillaud-Vivien et al., 2018, Example 2).

**SISOKR Excess-Risk.** We can now provide a bound on the excess-risk of SISOKR.

**Theorem 1** (SISOKR excess-risk bound)**.** *Let $\delta \in (0,1]$, $n \in \mathbb{N}$ such that $\lambda = n^{-1/(1+\gamma_{\mathcal{X}})} \geq \frac{9\kappa_{\mathcal{X}}{}^2}{n}\log(\frac{n}{\delta})$. Under Assumptions 1 to 4, with probability $1-\delta$ we have*

$$\mathbb{E}_x\left[\|\tilde{h}(x) - h^*(x)\|^2_{\mathcal{H}_\mathcal{Y}}\right]^{\frac{1}{2}} \leq S(n,\delta) + c_2 A^{\psi_\mathcal{X}}_{\rho_\mathcal{X}}(\widetilde{P}_X) + A^{\psi_\mathcal{Y}}_{\rho_\mathcal{Y}}(\widetilde{P}_Y)\,, \tag{10}$$

*where $S(n,\delta) = c_1 \log(4/\delta)\, n^{-\frac{1}{2(1+\gamma_\mathcal{X})}}$ and*

$$A^{\psi_\mathcal{Z}}_{\rho_\mathcal{Z}}(\widetilde{P}_Z) = \mathbb{E}_z\left[\|(\widetilde{\mathrm{P}}_\mathrm{Z} - I_{\mathcal{H}_\mathcal{Z}})\,\psi_\mathcal{Z}(z)\|^2_{\mathcal{H}_\mathcal{Z}}\right]^{\frac{1}{2}},$$

*with $c_1, c_2 > 0$ constants independent of $n$ and $\delta$.*

*Proof sketch.* The proof relies on a decomposition of the operator $\widetilde{H}$ such that $\tilde{h}(x) = \widetilde{H}\psi_\mathcal{X}(x)$, see (44). The first term in (10) corresponds to the non-sketched kernel Ridge regression error, and the second term to the input sketching error. The latter extends both the results of Ciliberto et al. (2020) to sketched estimators, and that of Rudi et al. (2015) to the vector vector-valued case. The third term, i.e., the output sketching error is specific to our framework and derives from the expression of $h^*$ and Jensen's inequality. □

The learning rate of the first term, i.e., the non-sketched kernel Ridge regression error, has been shown to be optimal under our set of assumptions in a minimax sense (Caponnetto and De Vito, 2007). The second and the third terms are approximation errors due to the sketching of the input and the output kernels, respectively. In particular, they write as *reconstruction errors* (Blanchard et al., 2007) associated to the random projection $\widetilde{\mathrm{P}}_\mathrm{X}$ and $\widetilde{\mathrm{P}}_\mathrm{Y}$ of the feature maps $\psi_\mathcal{X}$ and $\psi_\mathcal{Y}$ through the input and output marginal distributions.

**Sketching Reconstruction Error.** In Theorem 2, we give bounds on the sketching reconstruction error for the family of sub-Gaussian sketches, enlarging the scope of sketching distributions whose reconstruction error's bound is known —it was previously limited to uniform and approximate leverage scores sub-sampling sketches (Rudi et al., 2015). More generally, note that are admissible in our theoretical framework all sketching distributions for which concentration bounds on the induced empirical covariance operators can be derived, since quantity $A^{\psi_\mathcal{Z}}_{\rho_\mathcal{Z}}(\widetilde{P}_Z)$ is then easily controlled. We now recall the definition of sub-Gaussian sketches, and show how to bound their reconstruction error.

**Definition 1.** *A sub-Gaussian sketch $\mathrm{R}_\mathcal{Z} \in \mathbb{R}^{\mathrm{m}_\mathcal{Z} \times n}$ is composed of i.i.d. entries such that $\mathbb{E}\left[\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}\right] = 0$, $\mathbb{E}\left[{\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}}^2\right] = 1/m$ and $\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}$ is $\frac{{\nu_\mathcal{Z}}^2}{\mathrm{m}_\mathcal{Z}}$-sub-Gaussian, for all $1 \leq i \leq \mathrm{m}_\mathcal{Z}$ and $1 \leq j \leq n$, where $\nu_\mathcal{Z} \geq 1$.*

Recall that a standard normal r.v. is 1-sub-Gaussian. Moreover, by Hoeffding's lemma, any r.v. taking values in a bounded interval $[a,b]$ is $(b-a)^2/4$-sub-Gaussian. Hence, any sketch matrix composed of i.i.d. Gaussian or bounded r.v. is a sub-Gaussian sketch. Finally, note that $p$-sparsified sketches (El Ahmad et al., 2023) are sub-Gaussian with ${\nu_\mathcal{Z}}^2 = 1/p$, with $p \in ]0,1]$.

**Theorem 2** (sub-Gaussian sketching reconstruction error). *For $\delta \in (0, 1/e]$, $n \in \mathbb{N}$ sufficiently large such that $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \leq \|\,\mathrm{C}_{\mathcal{Z}}\,\|_{\mathrm{op}}/2$, then if*

$$\mathrm{m}_{\mathcal{Z}} \geq c_4 \max\left(\nu_{\mathcal{Z}}^2\, n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \nu_{\mathcal{Z}}^4 \log\left(1/\delta\right)\right), \tag{11}$$

*with probability $1-\delta$ we have*

$$\mathbb{E}_z\left[\|(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}})\,\psi_{\mathcal{Z}}(z)\|_{\mathcal{H}_{\mathcal{Z}}}^2\right] \leq c_3\, n^{-\frac{1-\gamma_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \tag{12}$$

*where $c_3, c_4 > 0$ are constants independents of $n, \mathrm{m}_{\mathcal{Z}}, \delta$.*

*Proof sketch.* The proof essentially consists in bounding the difference between the empirical covariance operator and its sketched counterpart in operator norm, see (89). The latter rewrites as a sum of sub-Gaussian random variables in a separable Hilbert space, and we invoke Koltchinskii and Lounici (2017, Theorem 9). □

Hence, depending on the regularity of the distribution (defined through our set of assumptions), one can obtain a small reconstruction error even with a small sketching size. For instance, if $\mu_{\mathcal{Z}} = \gamma_{\mathcal{Z}} = 1/3$, one obtains a reconstruction error of order $n^{-1/2}$ by using a sketching size of order $n^{1/2} \ll n$. As a limiting case, when $\mu_{\mathcal{Z}} = \gamma_{\mathcal{Z}} = 0$, one obtains a reconstruction error of order $n^{-1}$ when using a constant sketching size.

**Learning rates for SISOKR with sub-Gaussian sketches.** For the sake of presentation, we use $\lesssim$ to keep only the dependencies in $n, \delta, \nu, \gamma, \mu$. We note $a \vee b \coloneqq \max(a, b)$.

**Corollary 1** (SISOKR learning rates). *Consider the Assumptions of Theorems 1 and 2, that $\|\,\psi_{\mathcal{Y}}(y)\|_{\mathcal{H}_{\mathcal{Y}}} = \kappa_{\mathcal{Y}}$ for all $y \in \mathcal{Y}$, and $n \in \mathbb{N}$ such that $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}}/2$ for $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}\}$. Set*

$$\mathrm{m}_{\mathcal{Z}} \gtrsim \max\left(\nu_{\mathcal{Z}}^2\, n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \nu_{\mathcal{Z}}^4 \log\left(1/\delta\right)\right) \tag{13}$$

*for $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}\}$. Then with probability $1-\delta$*

$$\mathcal{R}(\tilde{f}) - \mathcal{R}(f^*) \lesssim \log\left(4/\delta\right) n^{-\frac{1-\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}}}{2(1+\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}})}}. \tag{14}$$

*Proof.* Using Theorems 1 and 2 to bound $A_{\rho_{\mathcal{X}}}^{\psi_{\mathcal{X}}}(\widetilde{P}_X)$ and $A_{\rho_{\mathcal{Y}}}^{\psi_{\mathcal{Y}}}(\widetilde{P}_Y)$ gives that with probability $1-\delta$ it holds $\mathbb{E}_x\left[\|\tilde{h}(x) - h^*(x)\|_{\mathcal{H}_{\mathcal{Y}}}^2\right]^{\frac{1}{2}} \lesssim \log\left(4/\delta\right) n^{-\frac{1-\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}}}{2(1+\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}})}}$. We then apply the comparison inequality (Ciliberto et al., 2020) to the loss $\Delta(y, y') = \|\,\psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(y')\|_{\mathcal{H}_{\mathcal{Y}}}^2$. □

This corollary shows that under strong enough regularity assumptions, the proposed estimators benefit from a close-to-optimal learning rate, even with small input and output sketching sizes. For instance, if $\mu_{\mathcal{X}} = \mu_{\mathcal{Y}} = \gamma_{\mathcal{X}} = \gamma_{\mathcal{Y}} = 1/3$, one obtains a learning rate of $\mathcal{O}(n^{-1/4})$, instead of the optimal rate of $\mathcal{O}(n^{-3/8})$ under the same assumptions, but only requiring sketching sizes $\mathrm{m}_{\mathcal{X}}, \mathrm{m}_{\mathcal{Y}}$ of order $n^{1/2} \ll n$. As a limiting case, when $\mu_{\mathcal{X}} = \mu_{\mathcal{Y}} = \gamma_{\mathcal{X}} = \gamma_{\mathcal{Y}} = 0$, one attains the optimal $\mathcal{O}(n^{-1/2})$ learning rate using constant sketching sizes.

# References


Alaoui, A. and Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Proc. of the 26th annual Conference on Learning Theory*, pages 185–209. PMLR.

Bakir, G., Hofmann, T., Smola, A. J., Schölkopf, B., and Taskar, B. (2007). *Predicting structured data*. The MIT Press.

Bauer, F., Pereverzev, S., and Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72.

Belanger, D. and McCallum, A. (2016). Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.

Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.-P., and Bach, F. (2020). Learning with differentiable perturbed optimizers.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., and Rieck, B. (2020). Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712.

Brault, R., Heinonen, M., and Buc, F. (2016). Random fourier features for operator-valued kernels. In *Asian Conference on Machine Learning*, pages 110–125. PMLR.

Brogat-Motte, L., Rudi, A., Brouard, C., Rousu, J., and d'Alché Buc, F. (2022). Vector-valued least-squares regression under output regularity assumptions. *Journal of Machine Learning Research*, 23(344):1–50.

Brouard, C., d'Alché-Buc, F., and Szafranski, M. (2011). Semi-supervised penalized output kernel regression for link prediction. In *International Conference on Machine Learning (ICML)*, pages 593–600.

Brouard, C., Shen, H., Dührkop, K., d'Alché-Buc, F., Böcker, S., and Rousu, J. (2016a). Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):28–36.

Brouard, C., Szafranski, M., and D'Alché-Buc, F. (2016b). Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels. *The Journal of Machine Learning Research*, 17(1):6105–6152.

Cabannes, V. A., Bach, F., and Rudi, A. (2021). Fast rates for structured prediction. In *conference on learning theory*, pages 823–865. PMLR.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61.

Chatalic, A., Carratino, L., De Vito, E., and Rosasco, L. (2022a). Mean nyström embeddings for adaptive compressive learning. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9869–9889. PMLR.

Chatalic, A., Schreuder, N., Rosasco, L., and Rudi, A. (2022b). Nyström kernel mean embeddings. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3006–3024. PMLR.

Ciliberto, C., Rosasco, L., and Rudi, A. (2016). A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 4412–4420.

Ciliberto, C., Rosasco, L., and Rudi, A. (2020). A general framework for consistent structured prediction with implicit loss embeddings. *J. Mach. Learn. Res.*, 21(98):1–67.

Collins, M. and Duffy, N. (2001). Convolution kernels for natural language. *Advances in neural information processing systems*, 14.

Cortes, C., Mohri, M., and Weston, J. (2005). A general regression technique for learning transductions. In *International Conference on Machine Learning (ICML)*, pages 153–160.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

Deshwal, A., Doppa, J. R., and Roth, D. (2019). Learning and inference for structured prediction: A unifying perspective. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*.

Drineas, P., Mahoney, M. W., and Cristianini, N. (2005). On the nyström method for approximating a gram matrix for improved kernel-based learning. *JMLR*, 6(12).

El Ahmad, T., Laforgue, P., and d'Alché Buc, F. (2023). Fast kernel methods for generic lipschitz losses via $p$-sparsified sketches. *Transactions on Machine Learning Research*.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:205–1.

Gärtner, T. (2008). *Kernels for Structured Data*, volume 72 of *Series in Machine Perception and Artificial Intelligence*. WorldScientific.

Geurts, P., Wehenkel, L., and d'Alché Buc, F. (2006). Kernelizing the output of tree-based methods. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 345–352, New York, NY, USA. Association for Computing Machinery.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. (2012). Conditional mean embeddings as regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1803–1810.

Gygli, M., Norouzi, M., and Angelova, A. (2017). Deep value networks learn to evaluate and iteratively refine structured outputs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1341–1351. JMLR.org.

Kalinke, F. and Szabó, Z. (2023). Nystr\" om $m$-hilbert-schmidt independence criterion. *arXiv preprint arXiv:2302.09930*.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110 – 133.

Kpotufe, S. and Sriperumbudur, B. K. (2020). Gaussian sketching yields a J-L lemma in RKHS. In Chiappa, S. and Calandra, R., editors, *AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 3928–3937. PMLR.

Lacotte, J. and Pilanci, M. (2020). Adaptive and oblivious randomized subspace methods for high-dimensional optimization: Sharp analysis and lower bounds. *arXiv preprint arXiv:2012.07054*.

LeCun, Y., Chopra, S., Ranzato, M., and Huang, F.-J. (2007). Energy-based models in document recognition and computer vision. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 1, pages 337–341. IEEE.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. (2021). Towards a unified analysis of random fourier features. *Journal of Machine Learning Research*, 22(108):1–51.

Lin, X. V., Singh, S., He, L., Taskar, B., and Zettlemoyer, L. (2014). Multi-label learning with posterior regularization.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Mahoney, M. W. et al. (2011). Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224.

Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. (2020). Kernel methods through the roof: Handling billions of points efficiently. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

Micchelli, C. A. and Pontil, M. (2005). On learning vector-valued functions. *Neural computation*, 17(1):177–204.

Niculae, V., Martins, A., Blondel, M., and Cardie, C. (2018). Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning (ICML)*, pages 3799–3808. PMLR.

Nowak, A., Bach, F., and Rudi, A. (2019). Sharp analysis of learning with discrete losses. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 1920–1929.

Nowozin, S. and Lampert, C. H. (2011). Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3-4):185–365.

Osokin, A., Bach, F. R., and Lacoste-Julien, S. (2017). On structured prediction theory with calibrated convex surrogate losses. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NeurIPS) 30:*, pages 302–313.

Pillaud-Vivien, L., Rudi, A., and Bach, F. (2018). Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31.

Rahimi, A. and Recht, B. (2007). Random features for large scale kernel machines. *NIPS*, 20:1177–1184.

Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110. Neural Networks and Kernel Methods for Structured Domains.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28.

Rudi, A., Canas, G. D., and Rosasco, L. (2013). On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075.

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 3215–3225.

Schymanski, E., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquiere, B., and Neumann, S. (2017). Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9:22.

Senkene, E. and Tempel'man, A. (1973). Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 13(4):665–670.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172.

Steinwart, I., Hush, D. R., Scovel, C., et al. (2009). Optimal rates for regularized least squares regression. In *COLT*, pages 79–93.

Sterge, N. and Sriperumbudur, B. K. (2022). Statistical optimality and computational efficiency of nystrom kernel pca. *Journal of Machine Learning Research*, 23(337):1–32.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484.

Weston, J., Chapelle, O., Vapnik, V., Elisseeff, A., and Schölkopf, B. (2003). Kernel dependency estimation. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 897–904. MIT Press.

Williams, C. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 682–688. MIT Press.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1-2):1–157.

Yang, T., Li, Y.-f., Mahdavi, M., Jin, R., and Zhou, Z.-H. (2012). Nyström method vs random fourier features: A theoretical and empirical comparison. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Yang, Y., Pilanci, M., Wainwright, M. J., et al. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023.

# Estimation of proportions under Open set Label Shift using Mahalanobis Projection

Bastien Dussap [1] & Gilles Blanchard [2] & Badr-Eddine Chérief-Abdellatif [3]

[1] *Laboratoire de Mathématiques d'Orsay, Inria, CNRS, Université Paris-Saclay, France, bastien.dussap@universite-paris-saclay.fr*
[2] *Laboratoire de Mathématiques d'Orsay, Inria, CNRS, Université Paris-Saclay, France, gilles.blanchard@universite-paris-saclay.fr*
[3] *CNRS, France, badr.eddine.cherief.abdellatif@gmail.com*

**Résumé.** Cette présentation aborde deux aspects clés de l'adaptation de domaine non supervisée : le Label Shift et son extension, l'open set label shift. Le Label Shift postule que la divergence entre les ensembles d'entraînement et de test réside uniquement dans la distribution des labels, tandis que l'open set label shift permet l'émergence de nouvelles classes dans la cible, tout en maintenant les distributions conditionnelles des classes invariantes. L'accent est mis sur l'estimation des proportions des labels dans l'échantillon test, un problème appelé quantification dans la littérature. S'appuyant sur des travaux antérieurs utilisant un embedding des classes via un classifier ou par des méthodes à noyaux, nous proposons d'utiliser la distance de Mahalanobis en estimant les matrices de covariances des différentes classes afin d'exploiter toute l'information disponible et non plus simplement les moyennes des embeddings. Dans cette présentation, nous mettons en avant deux théorèmes de consistance de cette méthode dans les deux contextes étudiés et appuyons nos résultats par des expériences numériques.

**Mots-clés.** Machine Learning, Label Shift, Quantification Learning, Adaptation de Domaine

**Abstract.** This presentation addresses two key aspects of unsupervised domain adaptation: Label Shift and its extension, the open set label shift. Label Shift posits that the discrepancy between training and testing sets lies solely in the distribution of labels, while open set label shift allows for the emergence of new classes in the target, while maintaining invariant conditional class distributions. The focus is on estimating label proportions in the test sample, a problem referred to as quantification in the literature. Building upon previous work utilizing class embedding through a classifier or kernel methods, we propose using the Mahalanobis distance by estimating the covariance matrices of different classes to leverage all available information, rather than just the means of embeddings. In this presentation, we highlight two consistency theorems of this method in the two studied contexts and support our findings by numerical experiments.

**Keywords.** Machine Learning, Label Shift, Quantification Learning, Adaptation Domain

# 1 Introduction

In this talk, we will focus on two particular instances of unsupervised domain adaptation: the estimation of the target label proportions under *Label Shift* [1, 8] and an extension proposed independently by Garg et al. [6] and Dussap et al. [2] named *Open set Label Shift*.
The **Label Shift** hypothesis states that the distributions of the training and test sets differ only in the marginal distribution of the label $p(y)$, while the conditional distributions of the covariates given the label $p(y|x)$, are assumed to be the same. For example, label shift occurs in infectious disease modelling where the covariates are the observed symptoms while the label is the underlying disease state. During an epidemic, the expected proportion of sick individuals is larger than usual, although the distribution of symptoms given the disease state remains unchanged.
**Open set label shift**, suppose that the label distribution changes arbitrarily and a new class emerges, but the class conditional distributions $p(y|x)$ remain domain invariant.

Several different objectives have been addressed in the literature under these assumptions, here we focused on on the estimation of the target label proportions. For label shift, this problem has many names such as *class ratio estimation* [7] or *posterior probability shift* [5] but the most commonly used is **quantification** [4]. To the best of our knowledge, only two papers have addressed the quantification problem under open set label shift [2, 6]. In this talk we continue the unification work we did and presented last year [2, 3].

Our main contribution is a variant of the method we proposed in our previous work where we embed the distribution using, for instance, a classifier or Kernel Mean Embedding [9]. We now rely on the Mahalanobis distance equipped with an estimate of the covariance matrices of the embeddings to exploit the information obtained from each point and not just the mean as we did. We obtain new theoretical results and show that this new method is still robust in the open set label shift setting.

Numerical experiments support our theory.

## 1.1 Notations

Formally, consider a covariate space $\mathcal{X}$, typically a subset of $\mathbb{R}^d$, and two label spaces $\mathcal{Y} = \{1, \dots, c\}$ and $\tilde{\mathcal{Y}} = \{0, \dots, c\}$. We define the *source* noted $\mathbb{P}$ and the *target* noted $\mathbb{Q}$, as different probability distributions over the covariate label space pair $\mathcal{X} \times \mathcal{Y}$ for the source and $\mathcal{X} \times \tilde{\mathcal{Y}}$ for the target.

The target label distribution is denoted $\alpha^* = (\alpha_i^*)_{i=0}^c$ while each class-$i$ conditional target distribution is denoted $\mathbb{Q}_i$. Similarly, the source label distribution is denoted $\beta^* = (\beta_i^*)_{i=1}^c$ while each class-$i$ conditional source distribution is denoted $\mathbb{P}_i$. We assume that the Open Set Label Shift hypothesis holds:

$$\mathbb{Q} = \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i + \alpha_0^* \mathbb{Q}_0 \qquad (\mathcal{OSLS})$$
$$\forall i = 1, \dots, c\,, \quad \mathbb{P}_i = \mathbb{Q}_i.$$

We recover the label shift setting when $\alpha_0^* = 0$:

$$\mathbb{Q} = \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i \qquad (\mathcal{LS})$$
$$\forall i = 1, \dots, c\,, \quad \mathbb{P}_i = \mathbb{Q}_i,$$

and in that case $\tilde{\mathcal{Y}} = \mathcal{Y}$.

A source dataset $\{(x_j, y_j)\}_{j\in[n]} \in (\mathcal{X} \times \mathcal{Y})^n$ and a target dataset $\{x_{n+j}\}_{j\in[m]} \in \mathcal{X}^m$ are given. All data points from the source (respectively the target) dataset are sampled independently from the source (resp. the target) domain. We have access to the source labels $y_j$ but not to the unobserved target labels. We denote by $\hat{\mathbb{P}}_i \coloneqq \sum_{j\in[n]:y_j=i} \delta_{x_j}(\cdot)/n_i$ the empirical conditional distribution of class $i$ from the source, where $n_i$ denotes the number of instances labelled $i$ in the source dataset. Note that $n_1 + \ldots + n_c = n$. We denote by $\tilde{\beta}$ the empirical proportions of each class in the source dataset, i.e. $\tilde{\beta}_i \coloneqq n_i/n$. Likewise, we denotes by $\hat{\mathbb{Q}} \coloneqq \sum_{j\in[m]} \delta_{x_{n+j}}(\cdot)/m$ the empirical distribution of the target.

Finally, for any function $\Phi : \mathcal{X} \to \mathcal{F}$, let's note $V = \left[\Phi(\hat{\mathbb{P}}_1), \cdots, \Phi(\hat{\mathbb{P}}_c)\right]$, $\boldsymbol{G} = V^T V$ the gram matrix of the source empirical embeddings and for any matrix (or operator) $M$, let's note $\boldsymbol{G}^M = V^T M^T M V$ the gram matrix associated with the Mahalanobis distance.

## 2 Distribution Feature Matching

In this section we present our previous work [3].

Let $\Phi : \mathcal{X} \to \mathcal{F}$ be a fixed feature mapping from $\mathcal{X}$ into a Hilbert space $\mathcal{F}$ (possibly $\mathcal{F} = \mathbb{R}^D$). We extend the mapping $\Phi$ to probability distributions on $\mathcal{X}$ by taking the expectation, i.e. $\Phi\colon \mathbb{P} \mapsto \mathbb{E}_{X\sim\mathbb{P}}[\Phi(X)] \in \mathcal{F}$. Thus it holds $\Phi(\hat{\mathbb{P}}_i) = n_i^{-1} \sum_{j\in[n]:y_j=i} \Phi(x_j)$, and similarly for $\Phi(\hat{\mathbb{Q}})$.
We call *Distribution Feature Matching* (DFM) any estimation procedure that can be formulated as the minimiser of the following problem:

$$\hat{\alpha} = \arg\min_{\alpha\in\Delta^c} \left\| \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right\|_{\mathcal{F}}^2 \qquad (\mathcal{P})$$

where $\Delta^c$ is the ($c$-1)-dimensional simplex.

This general formulation allows us to encompass other algorithms in the literature. Two in particular: *Kernel Mean Matching* [7] that rely on Kernel Mean Embedding and *BBSE* [8] that use the output of a classifier to embed the data.

## Theoretical guarantees

We make the following hypothesis on the mapping $\Phi$:

$$\sum_{i=1}^{c} \beta_i \Phi(\mathbb{P}_i) = 0 \iff \beta_i = 0 \ \forall i \tag{$\mathcal{A}_1$}$$

$$\|\Phi(x)\| \leq C \text{ for all } x. \tag{$\mathcal{A}_2$}$$

We can state the main theorem of [2]:

**Theorem 2.1.** *If the Label Shift hypothesis $\mathcal{LS}$ holds, and if the mapping $\Phi$ verifies the assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$*
*Then with probability greater than $1-\delta$ :*

$$\|\hat{\alpha} - \alpha^*\|_2 \leq \frac{2CR_{c/\delta}}{\sqrt{\lambda_{min}}} \left( \sqrt{\frac{\|w\|_1}{n}} + \frac{1}{\sqrt{m}} \right) \tag{1}$$

$$\leq \frac{2CR_{c/\delta}}{\sqrt{\lambda_{min}}} \left( \frac{1}{\sqrt{\min_i n_i}} + \frac{1}{\sqrt{m}} \right), \tag{2}$$

*where $R_x = 1 + \sqrt{2\log(2/x)}$, $w_i = \frac{\alpha_i^*}{\hat{\beta}_i}$ and $\lambda_{min} := \lambda_{min}(\boldsymbol{G}^M)$ is the smallest eigenvalue of $\boldsymbol{G}^M$.*

In other words, Theorem 2.1 states that under label shift and mild conditions on the embedding $\Phi$, the DFM procedure converges to the true proportions $\alpha^*$ with speed $\mathcal{O}\left((\min_i n_i)^{-1/2}\right)$ in the worst case and $\mathcal{O}\left((c/n)^{1/2}\right)$ in the best case scenario where the proportions of the source and the target haven't changed.

## 2.1 Open Set Label Shift

In our previous work, we proposed a new procedure called *soft*-DFM to deal with the Open Set Label Shift setting $\mathcal{OSLS}$. Since the proportions $\alpha^*$ we want to estimate no longer sum to one, the "hard" condition $\sum_i \alpha_i = 1$ in $\mathcal{P}$ is no longer needed:

$$\underset{\alpha \in \text{int}(\Delta^c)}{\arg\min} \left\| \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right\|_{\mathcal{F}}^2, \tag{$\mathcal{P}_2$}$$

where $\text{int}(\Delta^c) := \{x \in \mathbb{R}^c \colon x \geq 0, \ \sum x_i \leq 1\}$.

We had a theorem for *soft*-DFM procedure:

**Theorem 2.2.** *Denote by $\hat{\alpha}_{\text{soft}}$ the minimiser of the soft-DFM problem $\mathcal{P}_2$. Assume the Open Set Label Shift hypothesis ($\mathcal{OSLS}$) holds. If the mapping $\Phi$ verifies Assumptions ($\mathcal{A}_1$) and ($\mathcal{A}_2$). Then, with probability greater than $1-\delta$:*

$$\|\hat{\alpha}_{\text{soft}} - \alpha^*\|_2 \leq \frac{1}{\sqrt{\lambda_{min}}}\Big(3\epsilon_n + \varepsilon_m + \sqrt{2\alpha_0\ \epsilon_n\ \|\Phi(\mathbb{Q}_0)\|_{\mathcal{F}}} + \|\Pi_V(\Phi(\mathbb{Q}_0))\|_{\mathcal{F}}\Big),$$

*with:*

$$\epsilon_n = \frac{R_{\delta/\log c}}{\sqrt{\min_i n_i}}; \qquad \varepsilon_m = \frac{R_\delta}{\sqrt{m}};$$

*and $\Pi_V$ the orthogonal projection on $V$.*

Observe that the bound of theorem 2.2 shows robustness of DFM procedures against perturbations $\Phi(\mathbb{Q}_0)$ that are orthogonal to the source embeddings.
In particular if we use a classifier to embed the source distributions, the feature space has the same dimension as the number of sources, so under the condition ($\mathcal{A}_1$), $V$ will coincide with $E_1$, the affine space of vectors summing to one. Since any distribution will also be mapped to $E_1$, the orthogonal component will always be 0. Thus, we do not expect any particular robustness property for BBSE methods. On the other hand, if we use Kernel Mean Embedding with a translation invariant kernel i.e. $k(x,y) = \varphi(x-y)$, then for any distributions $\mathbb{P}, \mathbb{P}'$ it holds $\langle \Phi(\mathbb{P}), \Phi(\mathbb{P}')\rangle = \mathbb{E}_{(X,Y)\sim\mathbb{P}\otimes\mathbb{P}'}[\varphi(X-Y)]$. Thus, if $\varphi$ decays rapidly (e.g. Gaussian kernel), the feature mappings $\Phi(\mathbb{P})$ and $\Phi(\mathbb{P}')$ will be nearly orthogonal (have a scalar product close to 0) whenever the distributions $\mathbb{P}$ and $\mathbb{P}'$ are well separated. From this analysis, we expect that embedding the data using the Gaussian kernel will lead to a robust *soft*-DFM procedure for contamination distributions $\mathbb{Q}_0$ whose main mass is far from the source distributions.

# 3 Mahalanobis Distance

We now propose a new method where instead of minimising the $L_2$ distance between the embeddings as we did ($\mathcal{P}$), we minimise the Mahalanobis distance associated with an operator $M$:

$$\hat{\alpha} = \underset{\alpha\in\Delta^c}{\operatorname{argmin}} \left\| M\left(\sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}})\right)\right\|_{\mathcal{F}}^2 \qquad (\mathcal{MP})$$

To study the approximation error of our proposed estimator, we rely on a Bernstein inequality for Hilbert space-valued independent random variables due to Wolfer et al. [11], based on seminal work by Pinelis [10] and Yurinsky [12].

**Theorem 3.1** (Bernstein)**.** *Let $\mathcal{H}$ be a Hilbert space, and let $(z_i)$ be $n$ independent random variables (not necessarily identically distributed) with values in $\mathcal{H}$. Suppose that for each $i$, $\|z_i\| \leq C$ almost surely, where $C < \infty$. Denote $\overline{\Sigma} \coloneqq \frac{1}{n}\sum \Sigma_{z_i}$, where $\Sigma_{z_i}$ denotes the covariance matrix of $z_i$. Then for any $0 < \delta < 1$, with confidence $1-\delta$ it holds that,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} M(z_i - \mathbb{E}z_i) \right\| \leq \frac{2}{3}\sigma_1(M)\frac{CL_{1/\delta}}{n} + \sqrt{\frac{2L_{1/\delta}}{n} Tr(M\overline{\Sigma}M^\top)}, \tag{3}$$

*where* $L_x = \log(2x)$.

We can now state our main theorem: For any matrix $M$ (or any operator $M$ if $\mathcal{H}$ is an infinite dimensional space), under label shift we have

**Theorem 3.2.** *If the Label Shift hypothesis holds $\mathcal{LS}$ and if the mapping $\Phi$ verifies Assumptions $(\mathcal{A}_1)$, $(\mathcal{A}_2)$, then for any $\delta \in (0,1)$, with probability greater than $1-\delta$, the solution $\hat{\alpha}$ of $(\mathcal{MP})$ satisfies:*

$$\begin{aligned}\|\hat{\alpha} - \tilde{\alpha}\| \leq R_1(\delta, c)\frac{\|M\|_{\text{op}} C}{\sqrt{\lambda_{min}(\boldsymbol{G}^M)}}\left(\frac{\|w\|_1}{n} + \frac{1}{m}\right) \\ + R_2(\delta, c)\sqrt{\frac{Tr(M\Sigma_{\tilde{\alpha}}M^\top)}{\lambda_{min}(\boldsymbol{G}^M)}}\left(\sqrt{\frac{\|w\|_1}{n}} + \frac{1}{\sqrt{m}}\right),\end{aligned}$$

*with* $w = \frac{\tilde{\alpha}_i}{\hat{\beta}_i}$, $R_1(\delta, c) = 4/3\log(4c/\delta)$, $R_2 = 2\sqrt{2\log(4c/\delta)}$, $\Sigma_{\tilde{\alpha}} = \sum_{i=1}^{c} \tilde{\alpha}\Sigma_i$ *and* $\Sigma_i$ *the covariance matrix of* $\Phi(\mathbb{P}_i)$.

We have the same order of convergence as in theorem 2.1, but the constant before the term in $\mathcal{O}\left(n^{-1/2}\right)$ no longer depends on $C$ but on the trace of the covariance matrix.

The matrix $M$ which minimises the upper bound minimises :

$$\frac{Tr(M\Sigma_{\tilde{\alpha}}M^\top)}{\lambda_{\min}(\boldsymbol{G}^M)}. \tag{4}$$

The following theorem gives the close-form expression of this optimal $M$.

**Theorem 3.3.** *If the mapping $\Phi$ verifies Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_1)$. The matrix $M$ that minimise 4 verify:*

$$MM^\top = \Sigma_{\tilde{\alpha}}^{-1/2}\left(\Sigma_{\tilde{\alpha}}^{-1/2} V V^\top \Sigma_{\tilde{\alpha}}^{-1/2}\right)^{+} \Sigma_{\tilde{\alpha}}^{-1/2}, \tag{$\mathcal{M}$}$$

*Where the notation $()^+$ designate the Moore–Penrose inverse or Pseudo-inverse of a matrix.*

*4 is then equals to*

$$\text{Tr}\left(\left(\Sigma_{\tilde{\alpha}}^{-1/2} V V^\top \Sigma_{\tilde{\alpha}}^{-1/2}\right)^{+}\right) \tag{5}$$

Figure 1 shows the two bounds from theorem 2.1 and theorem 3.2 with respect to the number of points in the source and target. As we can see, this new bound is better than the previous one, especially for small numbers of points.

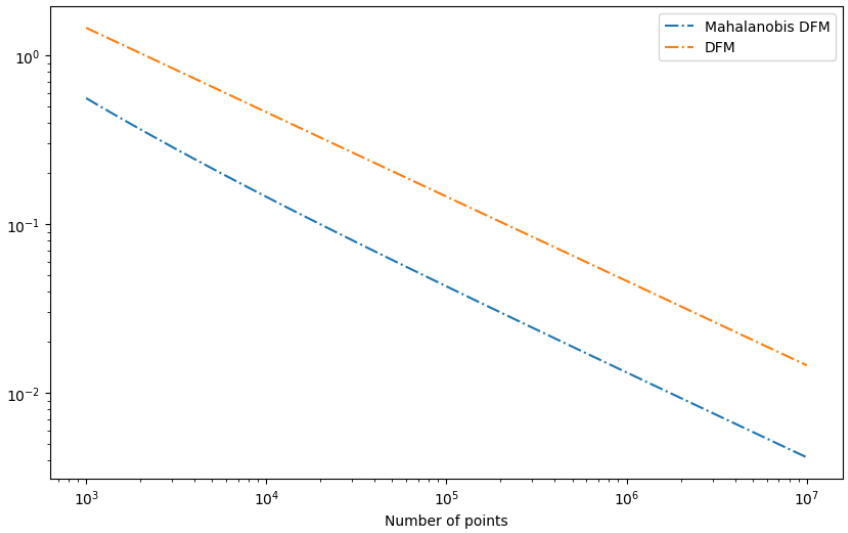



Figure 1: In blue the bound of theorem 3.2 and in yellow the bound of theorem 2.1. On the x-axis the number of points (in logarithmic scale) in the source and target, on the y-axis the value of both bounds (in logarithmic scale).

The default choice for $M$ would be to take $\tilde{M} = \Sigma_{\tilde{\alpha}}$. Without the constraint $\alpha \in \Delta^c$, the solution of $\mathcal{MP}$ for the optimal $M$ and $\tilde{M}$ are the same.

With the constraint however, the two solutions are different. For the optimal $M$ the solution is $\mathrm{Proj}_{\Delta^c}\Big(V^+ \mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big)$ while for $M = \Sigma^{-1/2}$ the solution is $V^+ \mathrm{Proj}_{\mathcal{C},\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)$, where $\mathcal{C}$ denote the convex hull of the empirical source embeddings, i.e. $\mathcal{C} \coloneqq \Big\{\sum \beta_i \Phi(\hat{\mathbb{P}}_i) \colon \beta \in \Delta^c\Big\}$. One can show that :

$$\mathrm{Proj}_{\Delta^c}\Big(V^+ \mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big) = V^+ \mathrm{Proj}_{\mathcal{C}}\Big(\mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big),$$
$$V^+ \mathrm{Proj}_{\mathcal{C},\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big) = V^+ \mathrm{Proj}_{\mathcal{C},\Sigma^{-1/2}}\Big(\mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big)\Big).$$

Put simply, the difference between the optimal choice of $M$ and the standard choice $M^*$ concerns only the renormalisation of the proportions given by $\mathrm{Proj}_{V,\Sigma^{-1/2}}\Big(\Phi(\hat{\mathbb{Q}})\Big)$. In the first case, the coordinates are projected orthogonally onto the simplex, while in the second case the coordinates are projected along the metric induced by $\Sigma^{-1/2}$. In particular, if $\mathrm{Proj}_{V,\Sigma^{-1/2}}\big(\Phi(\hat{\mathbb{Q}})\big) \in \mathcal{C}$ then the two solutions coincide.
This explains why in practice we don't see any difference between the two $M$.

## Open Set Label Shift and Mahalanobis

Similarly to $\mathcal{P}_2$, we propose *soft*-MDFM:

$$\arg\min_{\alpha \in \text{int}(\Delta^c)} \left\| M \left( \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right) \right\|_{\mathcal{F}}^2, \qquad (\mathcal{MP}_2)$$

Theorem 2.2 is still true when applied to *soft*-MDFM, except that the orthogonality condition that ensures robustness to the new class is now with respect to the Mahalanobis metric.

# 4 Experiments

In our experiment, the source consists of a list of $c$ Gaussian distributions: $\mathbb{P}_1, \cdots, \mathbb{P}_c$ in $\mathbb{R}^D$. The important parameter is $\rho$ and it controls how far the source is from the target. We tested with $c = 5$ and $c = 2$ classes, $D = 2, 5$ and 10, and we tested $\rho = 1$ we called it the "close setting" because the source and the target are close, and $\rho = 10$ the "far setting" where the source and the target are far from each other.
We refer to BBSE as the method that uses a classifier to embed the data, RFFM as the method that uses the Gaussian kernel, and MahalanobisRFFM as the method that uses the Gaussian kernel and a Mahalanobis distance with $M = \left(\hat{\Sigma} + \lambda I\right)$.

RFFM

| **Percentage of noise** $\epsilon$ | **Quantifier** | Number of classes = 5 | | |
|---|---|---|---|---|
| | | dim = 2 | dim = 5 | dim = 10 |
| 0.0 | BBSE | 4.11 ; 3.0 | 1.23 ; 3.0 | 0.97 ; 3.0 |
| 0.0 | RFFM | 1.75 ; 2.0 | 0.82 ; 2.0 | 0.84 ; 2.0 |
| 0.0 | MahalanobisRFFM | **1.55 ; 1.0** | **0.71 ; 1.0** | **0.71 ; 1.0** |
| 0.2 | BBSE | 26.90 ; 3.0 | 15.43 ; 3.0 | 12.42 ; 2.5 |
| 0.2 | RFFM | **16.20 ; 2.0** | 12.76 ; 2.0 | 12.22 ; 2.5 |
| 0.2 | MahalanobisRFFM | **17.38 ; 2.0** | **11.69 ; 1.0** | **10.94 ; 1.0** |
| 0.5 | BBSE | 52.43 ; 3.0 | 39.42 ; 3.0 | 31.95 ; 3.0 |
| 0.5 | RFFM | **30.70 ; 1.0** | 31.65 ; 2.0 | 30.51 ; 2.0 |
| 0.5 | MahalanobisRFFM | 33.98 ; 2.0 | **29.88 ; 1.0** | **27.88 ; 1.0** |
| 0.7 | BBSE | 67.25 ; 3.0 | 52.79 ; 3.0 | 44.55 ; 3.0 |
| 0.7 | RFFM | **45.76 ; 1.0** | 44.09 ; 2.0 | 41.21 ; 2.0 |
| 0.7 | MahalanobisRFFM | 52.71 ; 2.0 | **42.76 ; 1.0** | **39.50 ; 1.0** |

Table 1: **Gaussian Mixture: Comparison of BBSE, RFFM and MahalanobisRFFM when the new class is close**. The value before the semicolon is the geometric mean of the L2 error (multiplied by 100 for clarity) over 50 repetitions. Value after the semicolon is the median rank of a method relative to the other method in the same setting (same dimension, number of classes and percentage of noise). A bold value in a group is not significantly different from the best-performing method in the group (also in bold), as measured by a paired Wilcoxon test at $p < 0.01$.

| **Percentage of noise $\epsilon$** | **Quantifier** | Number of classes = 5 | | |
|---|---|---|---|---|
| | | dim = 2 | dim = 5 | dim = 10 |
| 0.0 | BBSE | 4.11 ; 3.0 | 1.23 ; 3.0 | 0.97 ; 3.0 |
| 0.0 | RFFM | 1.75 ; 2.0 | 0.82 ; 2.0 | 0.84 ; 2.0 |
| 0.0 | MahalanobisRFFM | **1.55 ; 1.0** | **0.71 ; 1.0** | **0.71 ; 1.0** |
| 0.2 | BBSE | 32.88 ; 3.0 | 23.96 ; 3.0 | 21.67 ; 3.0 |
| 0.2 | RFFM | **3.17 ; 1.5** | **1.85 ; 1.0** | **1.87 ; 1.0** |
| 0.2 | MahalanobisRFFM | **3.27 ; 1.5** | **1.92 ; 2.0** | **2.03 ; 2.0** |
| 0.5 | BBSE | 66.21 ; 3.0 | 60.31 ; 3.0 | 55.56 ; 3.0 |
| 0.5 | RFFM | **6.00 ; 1.0** | **4.42 ; 1.0** | **4.78 ; 1.0** |
| 0.5 | MahalanobisRFFM | **6.56 ; 2.0** | **4.50 ; 2.0** | **4.89 ; 2.0** |
| 0.7 | BBSE | 82.88 ; 3.0 | 77.70 ; 3.0 | 75.12 ; 3.0 |
| 0.7 | RFFM | **6.74 ; 1.0** | **5.64 ; 1.0** | **6.88 ; 1.0** |
| 0.7 | MahalanobisRFFM | **7.03 ; 2.0** | 5.94 ; 2.0 | **6.94 ; 2.0** |

Table 2: **Gaussian Mixture: Comparison of BBSE, RFFM and MahalanobisRFFM when the new class is far**. The value before the semicolon is the geometric mean of the L2 error (multiplied by 100 for clarity) over 50 repetitions. Value after the semicolon is the median rank of a method relative to the other method in the same setting (same dimension, number of classes and percentage of noise). A bold value in a group is not significantly different from the best-performing method in the group (also in bold), as measured by a paired Wilcoxon test at $p < 0.01$.

RFFM and its Mahlanobis version are both robust to noise that is far from the source, but not when it is close, whereas BBSE is never robust. This confirms our theoretical analysis. What we can also see is that even in the close setting, RFFM and MahalanobisRFFM both outperform BBSE, and MahalanobisRFFM is often slightly better than RFFM.

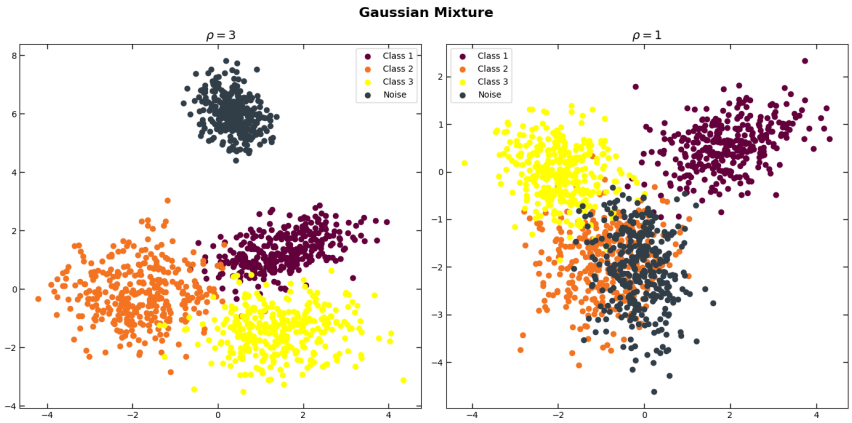



Figure 2: **Plot of the two settings in dimension 2 and with 3 classes.** On the left the new class in black is far from the sources in plum, orange and yellow. On the right the new class in black is close. Given of our theoretical analysis, we expect robustness for RFFM for the left setting and no robustness for the right setting.

# References


[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020.

[2] Bastien Dussap, Gilles Blanchard, and Badr-Eddine Chérief-Abdellatif. Label shift quantification with robustness guarantees via distribution feature matching. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 69–85. Springer Nature Switzerland, 2023.

[3] Bastien Dussap, Gilles Blanchard, and Badr-Eddine Chérief-Abdellatif. Label shift quantification with robustness guarantees via distribution feature matching. `https://bastiendussap.github.io/assets/files/slides/JdS2023.pdf`, 2023.

[4] Andrea Esuli, Alessandro Fabris, Alejandro Moreo, and Fabrizio Sebastiani. *Learning to Quantify*, volume 47. Springer Nature, 2023.

[5] Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the saerens-latinne-decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–34, 2020.

[6] Saurabh Garg, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems*, 35:22531–22546, 2022.

[7] Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538. PMLR, 2014.

[8] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[9] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[10] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[11] Geoffrey Wolfer and Pierre Alquier. Variance-aware estimation of kernel mean embedding. *arXiv preprint arXiv:2210.06672*, 2022.

[12] Vadim Yurinsky. *Sums and Gaussian vectors*. Springer, 2006.

# Régression Logistique One-hot pour la Classification

Baptiste Schall [1] & Rodolphe Anty [2] & Lionel Fillatre [3]

[1] *Université Côte d'Azur, France, bschall@i3s.unice.fr*
[2] *CHU Nice, Unité d'hépatologie, France, anty.r@chu-nice.fr*
[3] *Université Côte d'Azur, France, lionel.fillatre@i3s.unice.fr*

**Résumé**

Le classifieur de Bayes est très utilisé pour le traitement statistique des données. Lorsqu'on étudie des algorithmes d'apprentissage automatique tels que les réseaux de neurones, le classifieur de Bayes est souvent approché par une régression logistique, mais l'efficacité de cette approximation reste en partie incomprise. Dans cet article, nous proposons une régression logistique non-linéaire basée sur la discrétisation de descripteurs. Nous montrons que cette régression logistique peut parfaitement approcher le classifieur de Bayes naïf à condition d'appliquer un prétraitement spécifique détaillé dans cet article. En outre, grâce à cette méthode, chaque descripteur est associée à une fonction descriptive univariée dont les variations sont apprises de façon unique. Cette fonction descriptive nous permet d'interpréter la contribution d'un descripteur dans le processus de décision. Nous illustrons nos résultats théoriques à l'aide de données radiomiques.

**Mots-clés.** Classification binaire, Régression Logistique, Modèle additif, Optimalité Bayésienne, Discrétisation de descripteurs

**Abstract**

The Bayes classifier is widely used for statistical data analysis. When we exploit machine learning algorithms like neural networks, the Bayes classifier is often approximated with a logistic regression but the efficiency of this approximation is still questionable. In this paper, we propose a non-linear logistic regression based on features binning. We show that this logistic regression can perfectly approximate the naive Bayes classifier provided that the features are encoded with a specific preprocessing derived in this paper. Also, using this method each feature is associated with a feature function whose variations are learned uniquely. This feature function allows us to interpret the importance of the feature. We illustrate our theoretical results with radiomics data.

**Keywords.** Binary classification, Logistic Regression, Additive model, Bayes Optimality, Feature binning

# 1 Introduction

L'intelligence artificielle, dont les réseaux de neurones, est couramment utilisée dans le domaine du traitement du signal et des images [1]. Les méthodes d'apprentissage profond offrent de bonnes performances mais sont des "boîtes noires" dont les résultats sont souvent très difficiles à interpréter. La Régression Logistique (RL) est massivement utilisée en apprentissage profond en plus d'être très populaire dans le domaine médical.

Dans cet article, nous étudions la RL comme un modèle additif [2] avec décomposition sur une base de fonctions, i.e., la fonction de score de la RL est une somme de fonctions descriptives univariées constantes par morceaux. Chacune de ces fonctions nous permet d'interpréter l'impact de chaque descripteur sur la règle de décision. Cette modélisation est équivalente à la discrétisation des descripteurs et à leur encodage avec le très populaire encodage one-hot [3]. Pour garantir les performances de ce modèle, nous le comparons au Classifier Naïf discret de Bayes Discret (CNBD) [4–7]. Plusieurs articles comparent déjà le CNBD et la RL [8–11] en essayant notamment de créer des méthodes hybrides, par exemple en changeant la méthode d'apprentissage. Notre approche est semblable à [12] qui s'intéresse aux similitudes structurelles entre les modèles. En introduisant l'encodage one-hot [13] explicitement, nous simplifions l'utilisation de cette RL non-linéaire en plus d'un apport notable en terme d'interprétabilité. Nous explorons les deux aspects principaux de la comparaison entre RL et CNBD : l'erreur d'approximation et l'erreur d'estimation [3]. D'une part, nous montrons que l'erreur d'approximation entre la RL et le CNBD est nulle. D'autre part, nous montrons que l'erreur d'estimation liée à l'entraînement du RL nous empêche d'estimer l'ensemble des fonctions descriptives, mais que nous pouvons estimer avec précision leurs tendances.

Les contributions de cet article sont les suivantes. Tout d'abord, nous montrons qu'une RL dont les descripteurs sont one-hot encodés est équivalente à une RL dont la fonction de score est constante par morceaux ; chaque descripteur contribue à la fonction de score de façon non-linéaire via une fonction descriptive spécifique. Cette modélisation rend la RL one-hot encodée, appelée RLO, plus flexible. Ensuite, nous montrons que la RLO est équivalente au CNBD appliqué à des descripteurs discrétisés. Cette équivalence montre que la RLO est presque optimale sous les hypothèses habituelles requises par le CNBD. Troisièmement, nous montrons que les variations des fonctions descriptives peuvent être estimées de manière unique. Cette estimation est possible grâce à l'introduction d'un encodage spécifique, l'encodage one-hot suffisant. Enfin, nos expériences numériques établissent que la RLO est un bon compromis entre complexité et performance. Nous la comparons à des classifieurs plus complexes tels que le “Gradient-Boosting” et la “Random Forest” [2, 3].

La structure de l'article est la suivante. La section 2 étudie l'erreur d'approximation de la RLO. La section 3 introduit l'encodage suffisant et étudie l'estimation des fonctions descriptives. La section 4 illustre, à l'aide de données réelles, la pertinence de nos résultats théoriques. La section 5 conclut l'article.

## 2 Erreur d'approximation de la RL non-linéaire

Un échantillon $\underline{x}$ contient $d$ descripteurs tel que $\underline{x} = (x_1, \ldots, x_d)$. Les données sont normalisées telles que $x_i \in \Omega_i = \left[-1, 1\right]$ pour tout $i = 1, \ldots, d$. Le vecteur $\underline{x}$ est la réalisation du vecteur aléatoire $\underline{X} = (X_1, \ldots, X_d)$ composé de $d$ variables $X_i$.

La RL est un classifieur linéaire qui a comme objectif de modéliser la probabilité a posteriori $p(c|\underline{x})$ à l'aide de la fonction sigmoïde, $\sigma(t) = 1/(1+\exp(-t))$, appliquée à une fonction de score linéaire :

$$f_{\underline{\beta}}^{RL}(\underline{x}) = \sigma\left(\beta_0 + \beta_1 x_1 + ... + \beta_d x_d\right) = \sigma(h_{\underline{\beta}}^{RL}(\underline{x})). \tag{1}$$

Il est bien connu qu'étant donné une observation $\underline{x}$, la règle de décision Maximum A Posteriori (MAP) optimale [3] choisit la classe $c^*(\underline{x})$ qui maximise $p(c|\underline{x})$:

$$c^*(\underline{x}) = \underset{c \in \{0,1\}}{\arg\max}\, p(c|\underline{x}). \tag{2}$$

La RL est un modèle additif dans la mesure où la fonction de score $h^{RL}(\underline{x})$ dans (1) se réécrit

$$h_{\underline{\beta}}^{RL}(\underline{x}) = \beta_0 + f_{\beta_1}^{RL}(x_1) + ... + f_{\beta_d}^{RL}(x_d), \quad \text{où} \quad f_{\beta_i}^{RL} : x_i \mapsto \beta_i x_i \tag{3}$$

avec $\beta_i \in \mathbb{R}$. Nous proposons une variante de la RL, appelé la RLO ou RL avec encodage one-hot, avec des fonctions constantes par morceaux. Pour obtenir ce modèle, on partitionne le domaine de définition $\Omega_i$ des descripteurs $x_i$ en $m_i$ intervalles disjoints $I_{i,j}$ tels que $\Omega_i = \cup_{j=1}^{m_i} I_{i,j}$. Ainsi, la fonction constante par morceaux pour le descripteur $x_i$ est donnée par :

$$f_{\underline{\beta}_i}^{RLO} : x_i \mapsto \sum_{j=1}^{m_i} \beta_{i,j} \mathbb{1}\left\{x_i \in I_{i,j}\right\}, \tag{4}$$

où $\underline{\beta}_i = \left[\beta_{i,1}, ..., \beta_{i,m_i}\right] \in \mathbb{R}^{m_i}$ et $\mathbb{1}\{A\}$ est la fonction indicatrice qui est égale à 1 si l'événement $A$ est vrai et 0 sinon. De ce fait, la fonction descriptive $f_{\underline{\beta}_i}^{RLO}(x_i)$ sur le $j$ème intervalle $I_{i,j}$ du $i$ème descripteur est égale à $\beta_{i,j}$. Nous montrons dans la sous-section suivante que ce modèle constant par morceaux permet à la RLO d'approximer de manière précise la règle de décision optimale de Bayes naïf. Un descripteur $x_i$ one-hot encodée sera notée $\tilde{\underline{x}}_i \in \{0,1\}^{m_i}$ tel que :

$$\tilde{\underline{x}}_i = \left[\mathbb{1}\left\{x_i \in I_{i,1}\right\}, ..., \mathbb{1}\left\{x_i \in I_{i,m_i}\right\}\right]. \tag{5}$$

Quand $x_i \in I_{i,j}$, tous les éléments dans le vecteur $\tilde{\underline{x}}_i$ sont des zéros sauf le $j$ème On peut donc réécrire avec l'encodage one-hot :

$$f_{\underline{\beta}_i}^{RLO}(x_i) = \underline{\beta}_i^T \tilde{\underline{x}}_i, \tag{6}$$

où $\underline{x}^T$ est la transposée de $\underline{x}$.

## 2.1 Optimalité Bayésienne

Avec (6), la fonction de score de la RLO devient :

$$h_{\underline{\beta}}^{RLO}(\underline{x}) = \beta_0 + \sum_{i=1}^{d} \underline{\beta}_i^T \tilde{\underline{x}}_i. \tag{7}$$

Grâce à cette nouvelle écriture de la fonction de score de la RLO, nous pouvons montrer qu'elle correspond au Classificateur Naïf de Bayes (CNB). Il est bien connu que le CNB est le classificateur optimal lorsque les variables $X_i$ sont indépendantes conditionnellement à la classe $c$ [3]. La fonction de décision du CNB $f^{CNB}(\underline{x}) \in \{0, 1\}$ est

$$f^{CNB} = \underset{c\in\{0,1\}}{\arg\max}\, \mathbb{P}(C = c) \prod_{i=1}^{d} \mathbb{P}_c(X_i = x_i), \tag{8}$$

où $\mathbb{P}(C = c)$ est la probabilité a priori de la classe $c$ et $\mathbb{P}_c(X_i = x_i)$ est la probabilité conditionnelle de $X_i$ étant donné la classe $c$. Pour établir l'équivalence avec la RLO, nous devons approximer la probabilité $\mathbb{P}_c(X_i = x_i)$ sur la partition $\Omega_i = \cup_{j=1}^{m_i} I_{i,j}$. En d'autres termes, nous devons considérer le CNB discrétisé (CNBD) $f^{CNBD}$ donné par

$$f^{CNBD} = \underset{c\in\{0,1\}}{\arg\max}\, \mathbb{P}(C = c) \prod_{i=1}^{d} \prod_{j=1}^{m_i} \mathbb{P}_c(X_i \in I_{i,j})^{\mathbb{1}\{x_i \in I_{i,j}\}}. \tag{9}$$

Lorsque la taille maximale des intervalles $I_{i,j}$ est faible, l'écart entre le CNB et le CNBD est presque négligeable. On peut réécrire le CNB comme $f^{CNBD}(\underline{x}) = \mathbb{1}\{h^{CNBD}(\underline{x}) > 0\}$ où

$$h^{CNBD}(\underline{x}) = \alpha_0 + \sum_{i=1}^{d} \sum_{j=1}^{m_i} \alpha_{i,j} \mathbb{1}\{x_i \in I_{i,j}\}. \tag{10}$$

Les coefficients $\alpha_0$ et $\alpha_{i,j}$ sont donnés par :

$$\alpha_0 = \ln \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)}, \quad \alpha_{i,j} = \ln \frac{\mathbb{P}_1(X_i \in I_{i,j})}{\mathbb{P}_0(X_i \in I_{i,j})}. \tag{11}$$

Avec l'encodage one-hot $\tilde{\underline{x}}_i$, on peut réécrire $h^{CNBD}$ tel que :

$$h^{CNBD}(\underline{x}) = \alpha_0 + \sum_{i=1}^{d} \underline{\alpha}_i^T \tilde{\underline{x}}_i, \tag{12}$$

où $\underline{\alpha}_i = \left[\alpha_{i,1}, ..., \alpha_{i,m_i}\right]$. Nous pouvons remarquer que (7) coïncide avec (12) à l'exception de leurs coefficients qui sont différents : les coefficients CNBD sont déduits des distributions de probabilité discrétisées, tandis que les coefficients RLO sont appris à partir de l'ensemble de données d'apprentissage. Par conséquent, le modèle RLO est structurellement optimal par rapport au modèle CNBD. Ses coefficients peuvent être interprétés comme des rapports de vraisemblance optimaux (11). L'erreur d'approximation est donc nulle. Toutefois, comme le montre la section suivante, l'erreur d'estimation n'est pas négligeable. Il est essentiel de s'assurer que les paramètres du RLO peuvent être estimés de manière fiable.

# 3 Erreur d'estimation

## 3.1 Non unicité de l'estimation

Le modèle RLO est estimé avec l'entropie croisée binaire habituelle $\mathcal{L}(\underline{\beta})$ :

$$\mathcal{L}(\underline{\beta}) = -\sum_{i=1}^{N} c_i \ln\big(\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\big) + (1-c_i)\ln\big(1-\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\big) \tag{13}$$

où $\underline{\tilde{x}}_i$ désigne l'encodage (5) de $\underline{x}_i$. Soit $\underline{\hat{\beta}}$ un estimateur qui minimise $\mathcal{L}(\underline{\beta})$. Nous nous attendons bien sûr à ce que $\underline{\hat{\beta}}$ soit proche du vecteur de paramètres inconnu $\underline{\alpha}$. La RLO (12) est entièrement caractérisée par l'ensemble fini des scores $\sigma(\underline{\hat{\beta}}^\top \underline{\tilde{x}}_i)$ calculés sur l'ensemble du jeu de données $\mathcal{D}$. Par conséquent, la RLO est définie de manière unique par le vecteur de score $\tilde{X}\underline{\hat{\beta}}$ où $\tilde{X}$ est la matrice des données d'entrée (un échantillon par ligne) :

$$\tilde{X} = \left[\underline{\tilde{x}}_1^\top, \ldots, \underline{\tilde{x}}_N^\top\right]^\top \in \mathbb{R}^{N\times m}. \tag{14}$$

Le hessien $\nabla^2\mathcal{L}(\underline{\beta})$ de $\mathcal{L}(\underline{\beta})$ peut être réécrit :

$$\nabla^2\mathcal{L}(\underline{\beta}) = \tilde{X}^\top D(\underline{\beta})\tilde{X}, \tag{15}$$

$$D(\underline{\beta}) = \operatorname{diag}\left(\left[\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\left(1-\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\right)\right]\right), \tag{16}$$

où $\operatorname{diag}(\underline{u})$ désigne la matrice diagonale avec comme diagonale le vecteur $\underline{u}$. Puisque $D(\underline{\beta})$ est diagonale avec tous les éléments diagonaux non nuls, le rang de $\nabla^2\mathcal{L}(\underline{\beta})$ est égal à celui de la matrice de données $\tilde{X}$. On en déduit le lemme suivant.

**Lemme 1.** *La hessienne $\nabla^2\mathcal{L}(\underline{\beta})$ est semi-définie positive de rang $m-d$. La fonction $\mathcal{L}(\underline{\beta})$ n'est pas strictement convexe et l'estimateur $\underline{\hat{\beta}}$, s'il existe, n'est pas unique.*

## 3.2 Encodage Suffisant

Afin de garantir l'unicité de l'estimation remise en cause par le lemme 1, la proposition suivante introduit un nouveau encodage, appelé encodage suffisant, qui garantit la convexité stricte de $\mathcal{L}(\underline{\beta})$.

**Proposition 1.** *Il existe une matrice de permutation $Q$ telle que la matrice $\tilde{X}$ et le vecteur $\underline{\beta}$ satisfassent :*

$$\tilde{X}Q = \left[\tilde{S} \;\middle|\; \tilde{R}\right], \quad Q^\top\underline{\beta} = \begin{bmatrix}\underline{\theta}^S \\ \underline{\theta}^R\end{bmatrix}, \tag{17}$$

*où $\tilde{S}$ est de taille $N\times(m-d)$, $\tilde{R}$ est de taille $N\times d$, $\underline{\theta}^S \in \mathbb{R}^{m-d}$ et $\underline{\theta}^R \in \mathbb{R}^d$. Les propriétés suivantes sont vérifiées :*

- *$\tilde{S}$ est une matrice de rang plein colonne $m-d$,*

- *Il y a une matrice $L$ de taille $(m-d)\times d$, de rang plein colonne $d$, telle que $\tilde{R}=\tilde{S}L$.*
- *Les scores $\tilde{X}\underline{\beta}$ sur $\mathcal{D}$ sont préservés, i.e*

$$\tilde{X}\underline{\beta}=\tilde{S}(\underline{\theta}^S+L\underline{\theta}^R)=\tilde{S}\underline{\theta}, \tag{18}$$

*avec $\underline{\theta}=\underline{\theta}^S+L\underline{\theta}^R$.*

La proposition 1 introduit une nouvelle matrice d'encodage $\tilde{S}$ déduite de l'encodage initial $\tilde{X}$ qui est détaillé dans la sous-section 3.3. Ce nouvel encodage définit une nouvelle RL, appelé RLOS et notée $h_{\underline{\theta}}^{RLOS}(\underline{x})$.

**Lemme 2.** *Soit $\mathcal{L}(\underline{\theta})$ la fonction de perte à minimiser pour estimer $h_{\underline{\theta}}^{RLOS}(\underline{x})$. Le hessien $\nabla^2\mathcal{L}(\underline{\theta})\in\mathbb{R}^{(m-d)\times(m-d)}$ de $\mathcal{L}(\underline{\theta})$ est défini positif de rang $m-d$. La fonction $\mathcal{L}(\underline{\theta})$ est donc strictement convexe et l'estimateur $\hat{\underline{\theta}}$ qui minimise $\mathcal{L}(\underline{\theta})$, s'il existe, est unique.*

Grâce au lemme 2, nous avons la garantie d'obtenir une estimation unique $\hat{\underline{\theta}}$. Il s'agit d'un avantage significatif par rapport aux méthodes alternatives qui sont souvent utilisées pour obtenir une estimation raisonnable, comme une initialisation aléatoire pertinente de l'optimisation RL ou une régularisation de la fonction de perte (nous obtenons une estimation biaisée). Même si l'estimation de $\underline{\theta}$ est maintenant facilitée, il est important de reconstruire le vecteur $\underline{\beta}$ afin de trouver une interprétation de la RLOS en termes de fonctions descriptives constantes par morceaux (4). La proposition suivante montre qu'il existe un nombre infini de vecteurs $\underline{\beta}$ qui forment un espace linéaire de dimension $d$.

**Proposition 2.** *Soit les matrices $Q$ et $L$ définies dans la Proposition 1. Soit $\hat{\underline{\theta}}$ l'estimateur obtenu par le lemme 2. Tout RLO $h_{\underline{\beta}}^{RLO}(\underline{x})$ à coefficients $\underline{\beta}=\underline{\beta}(\underline{\theta},\underline{\theta}^R)$ de la forme*

$$\underline{\beta}(\hat{\underline{\theta}},\underline{\theta}^R)=Q\left(\begin{bmatrix}\hat{\underline{\theta}}\\0\end{bmatrix}+\begin{bmatrix}-L\\I_d\end{bmatrix}\underline{\theta}^R\right)=Q\Big(b(\hat{\underline{\theta}})+A\underline{\theta}^R\Big), \tag{19}$$

*où $\underline{\theta}^R\in\mathbb{R}^d$ est choisi arbitrairement, satisfait*

$$h_{\underline{\beta}}^{RLO}(\underline{x})=h_{\hat{\underline{\theta}}}^{RLSO}(\underline{x}),\forall\underline{x}. \tag{20}$$

## 3.3 Fonction descriptive constante par morceaux

D'après le lemme 1, l'estimation de la RLO n'est pas unique. Par conséquent, nous devons d'abord apprendre la RLOS. Pour apprendre une RLOS, nous devons encoder chaque variable avec l'encodage spécifique, appelé encodage suffisant, donné par la matrice $\tilde{S}$ dans la proposition 1. Cet encodage plus compact ne conserve que $m-d$ bits de l'encodage original $\underline{\tilde{x}}_i$. Les bits supprimés sont choisis arbitrairement mais, selon (18), cela n'affecte pas le score final. Il existe un moyen simple d'obtenir un encodage suffisant. Soit $\underline{\tilde{x}}_i^s$ les descripteurs codés tels que

$$\underline{\tilde{x}}_i^s=\left[\mathbb{1}\left\{x_i\in I_{i,1}\right\},...,\mathbb{1}\left\{x_i\in I_{i,m_i-1}\right\}\right]. \tag{21}$$

Le vecteur $\tilde{\underline{x}}_i^s$ est le vecteur $\tilde{\underline{x}}_i$ sans sa dernière composante. Nous pouvons alors construire $\tilde{S}$ dont les lignes sont $\tilde{\underline{x}}_i^s$. De cet encodage, nous pouvons déduire les matrices $Q$, $\tilde{R}$ et $L$ dans la Proposition 1. Ensuite, nous pouvons apprendre les coefficients uniques de la RLOS. Comme indiqué dans la proposition 2, étant donné la RLOS, nous pouvons construire un nombre infini de RLO en choisissant différents $\underline{\theta}_R$. Toutes ces RLO ont le même score mais pas les mêmes fonctions descriptives. Choisissons deux vecteurs arbitraires $\underline{\theta}^{R_1}$ et $\underline{\theta}^{R_2}$ et calculons $\hat{\underline{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_1})$ and $\hat{\underline{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_2})$. Avec (19), on a

$$\hat{\underline{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_1}) - \hat{\underline{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_2}) = Q \begin{bmatrix} -L \\ I_d \end{bmatrix} \left(\underline{\theta}^{R_1} - \underline{\theta}^{R_2}\right). \tag{22}$$

Ensuite, en utilisant les valeurs appropriées de $Q$ et $L$, un bref calcul montre que

$$\hat{\underline{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_1}) - \hat{\underline{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_2}) = \underline{\theta}_i^{R_1} - \underline{\theta}_i^{R_2} = \delta_i, \tag{23}$$

en notant $\underline{\theta}^{R_k} = \left[\underline{\theta}_1^{R_k}, \ldots, \underline{\theta}_d^{R_k}\right]$. Par conséquent, on obtient immédiatement que

$$f^{RLO}_{\hat{\underline{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_1})}(x_i) - f^{RLO}_{\hat{\underline{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_2})}(x_i) = \delta_i, \ \forall x_i \in \Omega_i, \tag{24}$$

à partir de la définition des fonctions descriptives dans (4). Par conséquent, bien que les fonctions descriptives ne soient pas uniques, elles sont simplement liées les unes aux autres par un décalage constant $\delta_i$ comme nous pouvons le voir sur la figure 1. Ce décalage peut être différent pour chaque descripteur. Par conséquent, l'interprétation des coefficients RLO est relative et non absolue : on interprète les variations des fonctions descriptives pour évaluer l'impact d'une variable d'entrée. En outre, dans la figure 1, nous pouvons noter la non-linéarité des fonctions descriptives RLO par rapport aux fonctions descriptives linéaires d'une RL conventionnelle.

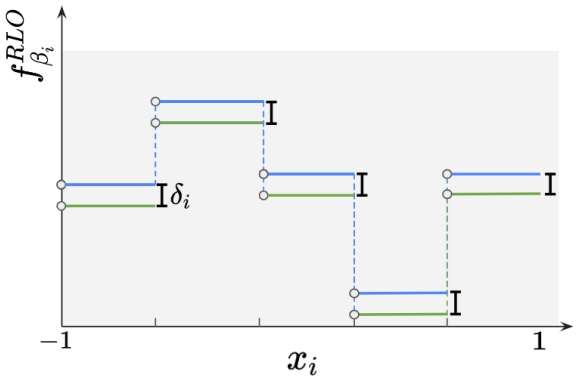



Figure 1: Exemple de fonctions descriptives de deux RLO, en vert et en bleu, provenant d'une même RLOS. Elles présentent les mêmes variations mais il y a un décalage vertical.

# 4 Expériences numériques

Nous avons testé notre méthode avec deux ensembles de données radiomiques créés à l'aide de la librairie PyRadiomic. La radiomique peut être définie comme un processus complexe visant à extraire des données numériques exploitables à partir d'images tomographiques. Les deux jeux de données proviennent du Mathematical Oncology Laboratory qui a publié en 2023 un jeu de données radiomiques massif [14] regroupant différents types de pathologie. Nous avons décidé de concentrer notre étude sur deux sous-ensembles de données qui en sont extraits. Le premier, RadioNsclc, traite les patients atteints d'un cancer du poumon non à petites cellules (Nsclc) et le second, RadioBreast, traite les patients atteints d'un cancer du sein. Pour les deux sous-ensembles de données, notre objectif est de prédire le temps de survie après la détection de la tumeur, qui est ici binaire (0 temps de survie court, 1 temps de survie long), déterminé en utilisant le temps de survie médian comme point critique. Ces ensembles de données contiennent respectivement 621 et 316 d'échantillons.

Il est commun d'utiliser des techniques de sélection de descripteurs en radiomique. Le problème de ces techniques est qu'elles introduisent un biais ; elles peuvent favoriser leur modèle sous-jacent. Par exemple, la sélection de descripteurs basée sur la RL avec une méthode LASSO [2] peut favoriser la sélection de variables qui interviennent de façon linéaire dans le modèle (au détriment de variables plus informatives qui interviendraient de façon non-linéaire). Par souci de simplicité, nous avons sélectionné manuellement les descripteurs pertinents pour mettre en évidence la non-linéarité en radiomique. Pour les deux ensembles de données, nous conservons une dizaine de descripteurs sur la centaine initiale.

Nous comparons d'abord la RL conventionnelle et la RLOS. Chaque modèle est entraîné à l'aide d'une validation croisée à 5 blocs. L'ensemble de données est discrétisé avec des intervalles de discrétisation variant de 5 à 10 en fonction du descripteur. La taille des intervalles est uniforme et déterminée par la règle de Sturges [15]. Les résultats sont présentés dans le tableau 1. La RL non-linéaire (RLOS) obtient de meilleures performances (+5-6%), ce qui prouve l'importance de la modélisation des effets non-linéaires. La figure 2 affiche la fonction descriptive RLO (4) pour le descripteur "Maximum 3D Diameter". Ce descripteur correspond à la plus grande distance euclidienne entre deux sommets du maillage représentant la tumeur en 3D. Comme indiqué dans la sous-section 3.3, ce sont les variations de la fonction descriptive RLO qui sont informatives, d'où la pertinence de n'étudier qu'une seule RLO. Des interprétations médicales pourraient découler de cette fonction descriptive.

| | **Train** | **Test** | **Train** | **Test** |
|---|---|---|---|---|
| **Méthodes** | **Précision moyenne (écart type)** | | | |
| **RL** | **69 (1)** | **64 (5)** | **68 (1)** | **64 (3)** |
| **RLOS** | **77 (2)** | **70 (4)** | **75 (2)** | **69 (3)** |
| **Données** | **RadioBreast** | | **RadioNsclc** | |

Table 1: Tableau des précisions et des écart-types sur l'ensemble d'entraînement "Train" et l'ensemble de test "Test" pour les ensembles de données RadioBreast et RadioNsclc pour différents modèles RL (la meilleure précision est surlignée).

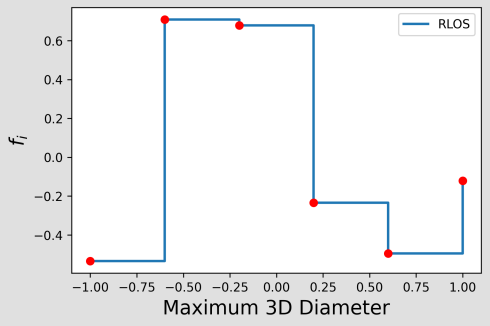



Figure 2: Fonction descriptive RLO du “Maximum 3D Diameter” pour RadioNsclc.

| | **Train** | **Test** | **Train** | **Test** |
|---|---|---|---|---|
| **Méthodes** | **Précision moyenne (écart type)** | | | |
| **RF** | **85 (1)** | **77 (3)** | **80 (1)** | **72 (3)** |
| **GB** | **84 (2)** | **76 (4)** | **86 (1)** | **76 (2)** |
| **Données** | **RadioBreast** | | **RadioNsclc** | |

Table 2: Tableau des précisions et des écart-types sur l’ensemble d’entraînement “Train” et l’ensemble de test “Test” pour les ensembles de données RadioBreast et RadioNsclc (la meilleure précision est surlignée).

Maintenant que nous avons vu que la RLOS est plus performante que la RL conventionnelle, nous pouvons comparer cette méthode avec des méthodes de classification avancées telles que le Gradient Boosting (GB) et les Random Forest (RF). Comme nous pouvons le voir dans le tableau 2, les modèles plus complexes obtiennent de meilleurs résultats. Cependant, ces méthodes souffrent d’un manque d’interprétabilité car leurs règles de décision sont très difficiles à comprendre. Par exemple, GB est basé sur un système de vote, dans notre cas un vote entre cent arbres sous-jacents, qui rend la règle de décision obscure. Le but de la RLOS n’est pas de surpasser ces méthodes mais de montrer qu’un modèle additif plutôt simple peut obtenir des résultats satisfaisants en plus d’un gain massif en interprétabilité. Enfin, la RLOS est très facile à configurer car les seuls hyper-paramètres sont liés à l’étape de discrétisation (nombre et largeur des intervalles). Pour entraîner les GB et RF, nous avons utilisé un algorithme dit “gridsearch” qui est à la fois fastidieux et chronophage.

# 5 Conclusion

Cet article montre que nous pouvons déduire une RL non-linéaire additive avec une performance et une interprétabilité accrues en représentant les descripteurs avec des vecteurs one-hot. Nous avons prouvé que cette RL one-hot est quasi-optimale au sens de Bayes. En outre, nous avons introduit l’encodage suffisant qui établit l’unicité des variations des fonctions descriptives associées aux variables d’entrées. Nos futurs travaux porteront sur la manière de prendre en compte les interactions entre les descripteurs.

# References


[1] Erik Meijering et al., "Deep learning in biological image and signal processing," *IEEE Signal Processing Magazine*, vol. 39, no. 2, pp. 24–26, 2022.

[2] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[3] Kevin P Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.

[4] Gherardo Varando et al., "Decision boundary for discrete bayesian network classifiers," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2725–2749, 2015.

[5] Daniel Berend and Aryeh Kontorovich, "A finite sample analysis of the naive bayes classifier," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1519–1545, jan 2015.

[6] Concha Bielza and Pedro Larranaga, "Discrete bayesian network classifiers: A survey," *ACM Computing Surveys*, vol. 47, pp. 1–43, 07 2014.

[7] Harry Zhang, "The optimality of naive bayes," in *Proceedings of the Seventeenth International FLAIRS Conference*, 2004.

[8] Andrew Ng and Michael Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *NIPS*, vol. 14, 2001.

[9] Chenyu Zheng et al., "Revisiting discriminative vs. generative classifiers: Theory and implications," *arXiv preprint arXiv:2302.02334*, 2023.

[10] P. Charan Kumar and B. T. Geetha, "Efficient removal of real time rain streaks from a image using novel naive bayes (NB) compare over linear regression (LR) with improved accuracy," in *International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 2023, pp. 1–6.

[11] Tapan Kumar Bhowmik, "Naive bayes vs logistic regression: Theory, implementation and experimental validation," *Inteligencia Artificial*, vol. 18, no. 56, pp. 14–30, Dec. 2015.

[12] Teemu Roos et al., "On discriminative bayesian network classifiers and logistic regression," *Machine Learning*, vol. 59, pp. 267–296, 2005.

[13] Samet Oymak, Mehrdad Mahdavi, and Jiasi Chen, "Learning feature nonlinearities with regularized binned regression," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1452–1456.

[14] Beatriz Ocaña-Tienda et al., "A comprehensive dataset of annotated brain metastasis MR images with clinical and radiomic data," *Scientific Data*, vol. 10, no. 1, pp. 208, 2023.

[15] David W Scott, "Sturges' rule," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, pp. 303–306, 2009.

# peerannot: A framework for label aggregation in crowdsourced datasets

Axel Dubar[1] & Tanguy Lefort[2] & Joseph Salmon[3]

[1] *Univ. Montpellier, IMAG, France, axel.dubar@umontpellier.fr*
[2] *Univ. Montpellier, CNRS, IMAG, Inria, LIRMM, France, tanguy.lefort@umontpellier.fr*
[3] *Univ. Montpellier, CNRS, IMAG, IUF, France, joseph.salmon@umontpellier.fr*

**Résumé.** Ce travail présente `peerannot`, une librairie de classification de données dont les étiquettes sont générées par production participative. Elle est écrite en Python et permet d'établir une comparaison des méthodes de classification par agrégation avec d'autres librairies de référence.

**Mots-clés.** Statistique Computationnelle, Classification et modèles de mélange

**Abstract.** This work presents `peerannot`, an image data classification library of image data whose labels are generated by crowdsourcing. It is written in Python and allows a comparison of aggregation classification methods with other reference libraries.

**Keywords.** Computational statistics, Classification, Clustering, mixture models.

## 1 Introduction

Crowdsourcing is a way of building large datasets faster, cheaper, and more easily than relying on experts. It consists of leveraging multiple workers to annotate data samples, where each worker considers a subsample of the dataset. It is a methodology that has been gaining in importance for the last two decades. Nowadays, many datasets (ImageNet[1], LabelMe[2], ...) have been collected through crowdsourcing. Yet, the use of multiple and often untrained workers can lead to wrong answers and disagreement among the collected labels. This is why a cleaning step is needed to identify an estimate of the correct label (hard label case) or a probability vector of the correct label (soft label case). Crowdsourced datasets pose significant problems for which the `peerannot` [3] library proposes a framework to handle this. In this work, we focus on the following: How can we effectively aggregate multiple labels into a single label from crowdsourced tasks?

## 2 `peerannot`: a framework for crowdsourced datasets

We introduce `peerannot`, `https://github.com/peerannot/peerannot`, an open-source Python library developed to address crowdsourced dataset annotation problems. Its main objective

is to provide a standardized library to ensure reproducibility and accessibility. We focus here on the aggregation methods, but `peerannot` also provides identification methods that allow the user to detect ambiguous tasks and possibly prune them from the dataset before training a classifier. Deep learning strategies (such as the CoNAL [4] or CrowdLayer [5]) are also available in the toolbox and can be used to classify images directly from the crowdsourced dataset. Finally, it provides a simulation module to generate crowdsourced datasets for testing purposes.

## 2.1 Notation

Before presenting the aggregation methods, we define some notation. A dataset $\mathcal{D} = (x_i, y_i^\star)_{i=1}^{n_{\text{task}}}$ is composed of $n_{\text{task}}$ tasks $x_i \in \mathcal{X}$ (the feature space) with unknown true labels $y_i^\star \in [K] = \{1, \dots, K\}$, with $K$ the number of classes. The indicator function is denoted by $\mathbb{1}(.)$ We use the $i$ index notation to range over the different tasks and the $j$ index notation for the workers. As the true label of the task $x_i$ is denoted by $y_i^\star$. The answer of the worker $w_j$ to the task $x_i$ is denoted by $y_i^{(j)}$ and the aggregated label from these answers is denoted by $\hat{y}_i$. We define the set of workers answering the task $x_i$:

$$\mathcal{A}(x_i) = \{j \in [n_{worker}] : w_j \text{ answered } x_i\} \ .$$

## 2.2 Aggregation methods

### 2.2.1 Peerannot methods

Focusing mainly on aggregation methods, we present the available strategies in `peerannot`.

- Majority vote (MV)
  This is the most basic model, it takes the label with the highest number of votes. It can be defined as:

$$\hat{y}_i^{\text{MV}} = \arg\max_{k \in [K]} \sum_{j \in \mathcal{A}(x_i)} \mathbb{1}_{\{y_i^{(j)} = k\}} \ .$$

- Naive Soft (NS)
  The Naive Soft model keeps the same intention as the Majority Vote but produces a soft label. By doing so it keeps track of the ambiguity among the workers for a task, a piece of information possibly discarded by the MV.

$$\hat{y}_i^{\text{NS}} = \left( \frac{1}{|\mathcal{A}(x_i)|} \sum_{j \in \mathcal{A}(x_i)} \mathbb{1}_{\{y_i^{(j)} = k\}} \right)_{j \in [K]} \ .$$

- Dawid and Skene (DS)
  The first two models do not take the worker's abilities into account. The Dawid and

**Algorithm 1** DS (EM version)

**Input**: $\mathcal{D}$: crowdsourced dataset

**Output**: $(\hat{y}_i^{\text{DS}})_{i\in[n_{\texttt{task}}]} = (\hat{T}_{i,.})_{i\in[n_{\texttt{task}}]}$: estimated soft labels and $\{\hat{\pi}^{(j)}\}_{j\in[n_{\texttt{worker}}]}$: estimated confusion matrices

1: **Initialization:** $\forall i \in [n_{\texttt{task}}], \forall \ell \in [K],\ \hat{T}_{i,\ell} = \frac{1}{|\mathcal{A}(x_i)|}\sum_{j\in\mathcal{A}(x_i)} \mathbb{1}_{\{y_i^{(j)}=\ell\}}$

2: **while** Likelihood not converged **do**

3: Get $\hat{\pi}$ and $\hat{\rho}$ assuming $\hat{T}$s are known

4: $\forall(\ell,k) \in [K]^2,\ \hat{\pi}_{\ell,k}^{(j)} \leftarrow \frac{\sum_{i\in[n_{\texttt{task}}]} \hat{T}_{i,\ell}\cdot\mathbb{1}_{\{y_i^{(j)}=k\}}}{\sum_{k'\in[K]}\sum_{i'\in[n_{\texttt{task}}]} \hat{T}_{i',\ell}\cdot\mathbb{1}_{\{y_{i'}^{(j)}=k'\}}}$

5: $\forall \ell \in [K],\ \hat{\rho}_\ell \leftarrow \frac{1}{n_{\texttt{task}}}\sum_{i\in[n_{\texttt{task}}]} \hat{T}_{i,\ell}$

6: Estimate $\hat{T}$s knowing $\hat{\pi}$ and $\hat{\rho}$

7: $\forall(i,\ell), \in [n_{\texttt{task}}]\times[K], \hat{T}_{i\ell} \leftarrow \frac{\prod_{j\in\mathcal{A}(x_i)}\prod_{k\in[K]}\hat{\rho}_\ell\cdot\left(\hat{\pi}_{\ell,k}^{(j)}\right)^{\mathbb{1}_{\{y_i^{(j)}=k\}}}}{\sum_{\ell'\in[K]}\prod_{j'\in\mathcal{A}(x_i)}\prod_{k'\in[K]}\hat{\rho}_{\ell'}\cdot\left(\hat{\pi}_{\ell'k'}^{(j')}\right)^{\mathbb{1}_{\{y_i^{(j')}=k'\}}}}$

8: **end while**

Skene's (DS) model [6], assumes each worker $w_j$ answers independently from one another. It assigns to each worker a confusion matrix $\pi^{(j)} \in \mathbb{R}^{K\times K}$ that represents the ability of a worker to answer a task given its label denoted as $\pi_{k,.}^{(j)}$. It represents the probability of the worker $j$ answering a task labeled $k$ correctly.
The associated likelihood can be written as:

$$\underset{\rho,\pi,T}{\arg\max} \prod_{i\in[n_{\texttt{task}}]}\prod_{k\in[K]}\left[\rho_k \prod_{j\in[n_{\texttt{worker}}]}\prod_{\ell\in[K]}\left(\pi_{k,\ell}^{(j)}\right)^{\mathbb{1}_{\{y_i^{(j)}=\ell\}}}\right]^{T_{i,k}}.$$

With $\rho_k = \mathbb{P}(y_i^\star = k)$ the probability of tasks labeled $k$ to appear in the dataset and $T_{i,k} = \mathbb{1}_{\{y_i^\star=k\}}$ the vectors of label class indicators for each task the true label $y_i^\star$ is unknown. The estimation procedure associated relies on the Expectation-Maximization (EM) algorithm, as described in Algorithm 1.

- Variations of DS model (FDS and WDS)
  The DS model is one of the most studied models in the literature and variations have been proposed such as a Fast Dawid and Skene's (FDS) [7] or a Weighted Dawid and Skene's (WDS) model (described in [8]).

- Generative model of Labels, Abilities, and Difficulties (GLAD)
  Now that workers' abilities are taken into account, the GLAD model also takes the task difficulty into account in the model to recover the soft label $\hat{y}_i^{GLAD}$.
  Defining $\alpha_j \in \mathbb{R}$ as the worker ability and $\beta_i \in \mathbb{R}_*^+$ as the task's difficulty, the GLAD model can be defined as:

$$\forall k \in [K],\ \mathbb{P}(y_i^{(j)} = k | y_i^\star \neq k, \alpha_j, \beta_i) = \frac{1}{K-1}\left(1 - \frac{1}{1+\exp(-\alpha_j\beta_i)}\right).$$

**Algorithm 2** GLAD (EM version)

**Input**: $\mathcal{D}$: crowdsourced dataset

**Output**:$\alpha = \{\alpha_j\}_{j\in[n_{\texttt{worker}}]}$: worker abilities, $\beta = \{\beta_i\}_{i\in[n_{\texttt{task}}]}$: task difficulties, aggregated labels

1: **while** Likelihood not converged **do**
2: Estimate probability of $y_i^\star$
3: $\forall i \in [n_{\texttt{task}}],\ \mathbb{P}(y_i^\star|\{y_i^{(j)}\}_i, \alpha, \beta_i) \propto \mathbb{P}(y_i^\star)\prod_j \mathbb{P}(y_i^{(j)}|y_i^\star, \alpha_j, \beta_i)$
4: Maximization step
5: Maximize auxiliary function $Q(\alpha, \beta)$ in Eq. 1 *w.r.t.* $\alpha$ and $\beta$
6: **end while**

The auxiliary function for the binary GLAD model is:

$$Q(\alpha,\beta) = \mathbb{E}[\log \mathbb{P}(\{y_i^{(j)}\}_{i,j}, \{y_i^\star\}_i)] = \sum_i \mathbb{E}[\log \mathbb{P}(y_i^\star)] + \sum_{i,j} \mathbb{E}[\log \mathbb{P}(y_i^{(j)}|y_i^\star, \alpha_j, \beta_i)] \ . \tag{1}$$

### 2.2.2 Other methods

Other competitors we compared with include the following:

- MMSR [9]: The Matrix Mean-Subsequence-Reduced strategy considers the reliability of all workers as a vector $s \in \mathbb{R}^{n_{\texttt{worker}}}$. Each entry $s_j$ represents the reliability of the worker $j$. This strategy assumes that each worker answers independently. It also assumes that a worker is correct with probability $p_j \in [0, 1]$ and the worker's probability of being wrong is uniform across classes, *i.e.*:

$$\forall (i,j) \in [n_{\texttt{task}}] \times [n_{\texttt{worker}}],\ \begin{cases} \mathbb{P}(y_i^{(j)} = k) = p_j & \text{if } y_i^\star = k, \\ \mathbb{P}(y_i^{(j)} = k) = \frac{1-p_j}{K-1} & \text{if } y_i^\star \neq k \end{cases} .$$

The reliability of a worker is linked to its probability of answering correctly: $s_j = \frac{K}{K-1}p_j - \frac{1}{K-1}$. This reliability can be estimated by solving a rank-one matrix completion problem defined as:

$$\mathbb{E}\left[\frac{K}{K-1}C - \frac{1}{K-1}\mathbb{1}\mathbb{1}^\top\right] = ss^\top \ ,$$

where $C$ is the covariance matrix of the workers' answers. More precisely, given two workers $j, j' \in [n_{\texttt{worker}}]$, the covariance between them is $C_{j,j'} = \frac{1}{N_{j,j'}}\sum_{i=1}^{n_{\texttt{task}}} \mathbb{1}(y_i^{(j)} = y_i^{(j')})$, with $N_{j,j'}$ the number of tasks in common: $N_{j,j'} = |\{i \in [n_{\texttt{task}}]|j, j' \in \mathcal{A}(x_i)\}|$. The final label is a weighted majority vote:

$$\hat{y}_i^{\text{M-MSR}} = \text{WMV}(i, W) \quad \text{with} \quad W_{j,k} = \log \frac{(K-1)p_j}{1-p_j} \ , \tag{2}$$

where the form of the weights is derived from a maximum a posteriori formulation of the model, see [10, Corollary 9].

- WAWA [11]: This strategy, also known as the inter-rater agreement, weights each user by how much they agree with the MV labels on average. More formally, given a task $i$:

$$\hat{y}_i^{\text{WAWA}} = \text{WMV}\left(i, W\right), \quad \text{with} \quad W_{j,:} = \left( \frac{1}{|\{y_{i'}^{(j)}\}_{i'}|} \sum_{i'=1}^{n_{\text{task}}} \mathbb{1}\left(y_{i'}^{(j)} = \hat{y}_{i'}^{\text{MV}}\right) \right) \mathbf{1}_K \ . \tag{3}$$

where $\forall i \in [n_{\texttt{task}}], \quad \hat{y}_i = \text{WMV}(i, W) := \arg\max_{k \in [K]} \sum_{j \in \mathcal{A}(x_i)} W_{j,k} \mathbb{1}(y_i^{(j)} = k)$, and $W \in \mathbb{R}^{n_{\texttt{worker}} \times K}$ is the matrix assigning the weight of worker $j$ when answering class $k$. It allows us to instantiate a weight that can vary for each worker (but not per task) and it usually improves on the MV strategy.

# 3 `benchopt`: a benchmark framework for optimization

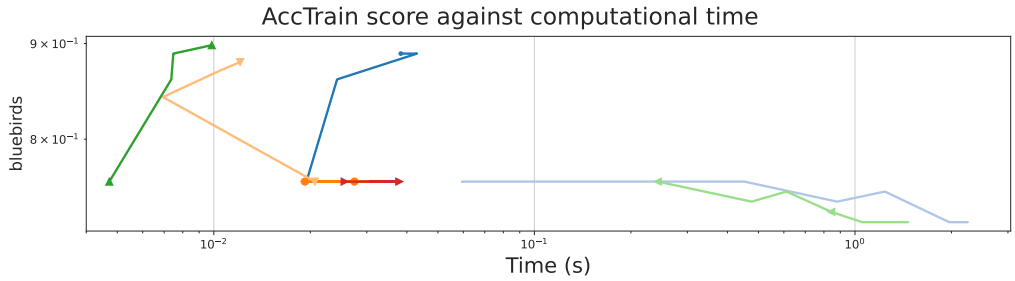
`benchopt` is an optimization benchmark library designed to facilitate comparing and reproducing optimization problems across different frameworks. It is an open-source library available at `https://github.com/benchopt/benchopt` that facilitates efficient and collaborative benchmarking by providing a standardized platform for researchers. With `benchopt`, users can easily explore, assess, and compare the performance of optimization algorithms on diverse problem sets. `benchopt` is designed for efficiency, enabling users to measure the performance of optimization algorithms by assessing the cumulated time taken to reach an optimum during the optimization steps. A list of already available implementation benchmarks such as OLS optimization or the LASSO optimization is available at `https://benchopt.github.io/available_benchmarks.html`.

# 4 Comparisons

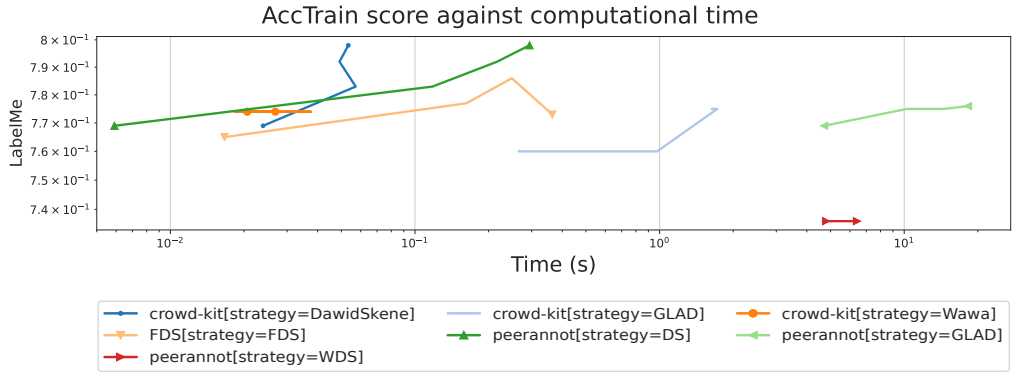
To compare different strategies and different implementations across libraries, a crowdsourcing benchmark has been implemented with the `benchopt` library, this benchmark can be found at `https://github.com/benchopt/benchmark_crowdsourcing`. For each strategy, we measure the cumulated time taken to reach an optimum in the accuracy metric. Each strategy is run 5 times until convergence. We measure the accuracy with the AccTrain metric, a metric defined as:

$$\text{AccTrain}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathbb{1}_{\{y_i = \arg\max_{k \in [K]} (\hat{y}_i)_k\}}.$$

It computes the number of correct predictions assuming that the ground truth is known, a condition that is not always guaranteed with crowdsourced datasets. It is important to note that some strategies such as the MV and NS are computed and do not need optimization steps so their presence in the benchmark is to be viewed differently. Their representation is not shown in the figures below but if necessary they would need to be viewed as points and not dotted lines. Another library focused on crowdsourcing has been implemented named `crowd-Kit` which can be used to perform comparisons with `peerannot`.




(a) Bluebirds dataset




(b) LabelMe dataset

Figure 1: Comparison of computational time (in seconds) for all aggregation strategies

To perform the measure of the computational time and accuracy of the strategies we run a benchmark on two datasets. The first is the Bluebirds dataset. A small dataset composed of 39 workers, 108 tasks, and $K = 2$ classes. The second is the LabelMe dataset [2]. Bigger than Bluebirds with 77 workers, 1000 tasks, and $K = 8$. The benchmark can be reproduced with the following command from the crowdsourcing `benchopt` branch:

```
benchopt run ./benchmark_crowdsourcing
```

As we can see in Figure 1a, the DS implementation from `peerannot` is the first to reach its convergence followed by the FDS, then the DS from `crowd-Kit`. The fact that the DS from `peerannot` is faster than the DS from `crowd-Kit` is not guaranteed for every dataset as it is not the case in Figure 1b. `crowd-Kit` implementation is based on aggregations using vanilla pandas which can be slower. Slowness can also be explained by initialised priors which can lead to a faster convergence.
Strategies not based on DS reach lower performance as it is a binary classification problem with a low number of workers. It can be noted that using confusion matrices, a worker's answers that are consistently wrong are still informative.

# 5 Conclusion

We have presented a new library to address crowdsourced dataset annotation problems and an extension of `benchopt` to allow users to compare different aggregation strategies. `peerannot` and `crowd-kit` libraries can both handle classification crowdsourced datasets. We presented their aggregation methods, however, other modules allowing the identification of poorly performing workers or hardest tasks are also proposed. On these tasks, both libraries differ, with `peerannot` proposing more diverse strategies for identification.

**Acknowledgments**: This work was supported in part by the French National Research Agency (ANR) through the grant ANR- 20-CHIA-0001-01 (Chaire IA CaMeLOt).

# References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR*. 2009.

[2] F Rodrigues, F Pereira, and B Ribeiro. "Gaussian process classification and active learning with multiple annotators". In: *ICML*. PMLR. 2014, pp. 433–441.

[3] T. Lefort, B. Charlier, A. Joly, and J. Salmon. "Peerannot: classification for crowdsourced image datasets with Python". In: *Computo* (2024). URL: `https://tanglef.github.io/computo_2023/`.

[4] Z Chu, J Ma, and H Wang. "Learning from Crowds by Modeling Common Confusions." In: *AAAI*. 2021, pp. 5832–5840.

[5] F Rodrigues and F Pereira. "Deep learning from crowds". In: *AAAI*. Vol. 32. 2018.

[6] AP Dawid and AM Skene. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *J. R. Stat. Soc. Ser. C. Appl. Stat.* 28.1 (1979), pp. 20–28. (Visited on 10/10/2021).

[7] V B Sinha, S Rao, and V N Balasubramanian. "Fast Dawid-Skene: A fast vote aggregation scheme for sentiment classification". In: *arXiv preprint arXiv:1803.02781* (2018).

[8] T Lefort, B Charlier, A Joly, and J Salmon. "Identify ambiguous tasks combining crowdsourced labels by weighting Areas Under the Margin". In: *arXiv* (2022).

[9] Q Ma and A Olshevsky. "Adversarial crowdsourcing through robust rank-one matrix completion". In: *NeurIPS*. Vol. 33. 2020, pp. 21841–21852.

[10] Hongwei Li and Bin Yu. "Error rate bounds and iterative weighted majority voting for crowdsourcing". In: *arXiv preprint arXiv:1411.4086* (2014).

[11] Appen Limited. *Calculating Worker Agreement with Aggregate (Wawa)*. 2021. URL: `https://success.appen.com/hc/en-us/articles/202703205-Calculating-Worker-Agreement-with-Aggregate-Wawa-`.

# Sensibilité des indices de qualité d'un classifieur probabiliste

Ndèye Awa Dieye[1,2], Ndèye Niang [1,2] & Giorgio Russolillo [1,2]

[1] *Laboratoire Cédric-Cnam, Paris*
[2] {*ndeye-awa.dieye, ndeye.niangkeita, giorgio.russolillo*}*@lecnam.net*

**Résumé.** Dans le domaine de la classification supervisée, évaluer les classifieurs probabilistes implique de regarder la discrimination, qui est la capacité à distinguer les classes, et la calibration, qui est la fiabilité de l'estimation des probabilités qui génèrent la variable de réponse. Dans ce papier, nous étudions la sensibilité de plusieurs mesures de qualité d'un classifieur probabiliste binaire via des simulations. Le but est de comprendre le comportement de ces mesures face à diverses formes de distribution des probabilités qui génèrent la variable de réponse et aux caractéristiques des écarts entre ces probabilités et leurs estimations.

**Mots-clés.** Classification supervisée, discrimination, calibration.

**Abstract.** In the field of supervised classification filed, assessing probabilistic classifiers involves examining discrimination, which is the ability to distinguish between classes, and calibration, which is the reliability of estimating the probabilities that generate the response variable. In this paper, we investigate the sensitivity of various quality measures for a binary probabilistic classifier through simulations. The objective is to comprehend the behavior of these measures in the presence of different shapes of the probability distribution generating the response variable, as well as the characteristics of deviations between these probabilities and their estimates.

**Keywords.** Supervised classification, discrimination, calibration.

# 1 Introduction

Les classifieurs probabilistes, largement utilisés dans divers domaines tels que la médecine, la finance et la reconnaissance d'images, estiment la probabilité $\pi$ d'un évènement pour chaque observation $i \in \{1, \ldots, n\}$. Dans le contexte de la classification binaire, les observations sont décrites par un ensemble de variables explicatives et une variable de réponse binaire $y \in \{0, 1\}$, qui indique l'occurrence de l'évènement. Ainsi, $\pi_i$ représente la vraie probabilité (inconnue en pratique) qui a généré la réponse $y_i$ pour la $i$ème observation. Celle-ci est à distinguer de la probabilité $p_i$ *estimée* par un classifieur binaire. Pour évaluer la qualité du classifieur, il faut mesurer la précision de l'estimation des probabilités $\pi_i$ et la qualité des prédictions des valeurs $y_i$ de la variable cible.

Plusieurs indices sont utilisés pour évaluer la qualité des classifieurs binaires. Ce travail présente une étude de la sensibilité de certains de ces indices. Nous définissons la sensibilité d'un indice comme son comportement face à des erreurs d'intensité différente dans

l'estimation des vraies probabilités. Via des simulations, on étudie cette sensibilité en fonction de 2 facteurs : la distribution des vraies probabilités $\pi_i$ et les caractéristiques des écarts (monotonie, symétrie, etc.) entre les probabilités estimées et les vraies probabilités.

Dans les sections suivantes, nous détaillons les mesures utilisées pour évaluer la performance des classifieurs binaires, nous présentons notre méthodologie expérimentale, puis nous discutons des résultats obtenus et de leurs implications en pratique.

# 2 Mesures de la performance d'un classifieur probabiliste

L'évaluation d'un classifieur se fait en examinant son pouvoir discriminant, sa calibration ou les deux à la fois. Le pouvoir discriminant est la capacité à bien classer les individus. La calibration mesure la fiabilité de l'estimation de la probabilité d'appartenance aux classes.

Dans la pratique, la capacité discriminante est communément évaluée par le taux d'erreur ainsi que de nombreux indices (taux de vrais positifs, taux de vrais négatifs, f1-score, etc.). Toutefois, ces indices dépendent du seuil de probabilité choisi pour faire le classement.

Cependant, d'autres mesures existent prenant en compte tous les seuils possibles comme l'aire sous la courbe ROC (AUC). La courbe ROC représente le taux de vrais positifs et le taux de vrais négatifs pour différents seuils de classement et l'AUC exprime la probabilité qu'un exemple positif sélectionné au hasard $x_j$ obtienne une probabilité plus élevée qu'un exemple négatif sélectionné au hasard $x_i$. Elle est donc une mesure de concordance. Soit $y^0$ l'ensemble des $n_0$ individus de classe 0 et $y^1$ l'ensemble des $n_1$ individus de classe 1, l'AUC est estimée comme (Calders & al. (2007)) :

$$AUC := \frac{1}{n_0 \cdot n_1} \left( \sum_{x_i \in y^0} \sum_{x_j \in y^1} \mathbb{1}[p_{x_i} < p_{x_j}] + \frac{1}{2} \sum_{x_i \in y^0} \sum_{x_j \in y^1} \mathbb{1}[p_{x_i} = p_{x_j}] \right)$$

où $\mathbb{1}[p_{x_i} < p_{x_j}]$ nous donne 1 si $p_{x_i} < p_{x_j}$ et 0 sinon. L'AUC varie dans $[0; 1]$.

Pour ce qui concerne l'évaluation de la calibration, on distingue la calibration *in the large* (calibration en moyenne par Van Calster & al. (2016)) et la calibration *in the small*. La première est issue de la comparaison du taux observé $\overline{y}$ de l'évènement et la moyenne $\overline{p}$ des probabilités estimées. Dans ce travail, nous choisissons de la calculer comme le rapport entre ces deux quantités :

$$Mean\_calib = \overline{y}/\overline{p}$$

Cet indicateur est égal à 1 lorsque le modèle est parfaitement calibré en moyenne. Lorsqu'il est supérieur à 1, il y a sous-estimation et surestimation dans le cas contraire.

Afin de mesurer la calibration *in the small*, on ordonne les individus selon les probabilités *estimées*, on les partitionne en $G$ groupes de taille $n_g$ et on compare le taux observé

d'évènement $\overline{y}_g$ et la moyenne des probabilités estimées $\overline{p}_g$ dans chaque groupe $g \in \{1, \ldots, G\}$. Dans ce travail, les groupes sont définis par les déciles de la distribution des probabilités estimées.

La calibration *in the small* peut être graphiquement évaluée à l'aide du diagramme de fiabilité (Figure 1). Il représente les taux d'évènement $(\overline{y}_g)$ en fonction des moyennes des probabilités prédites $(\overline{p}_g)$ de chaque groupe. Si le modèle est parfaitement calibré, alors les points seront tous sur la première bissectrice. Toute déviation, pour un groupe, de cette situation parfaite traduit une mauvaise calibration pour un certain niveau de probabilité (Pepe & al. (2013)).

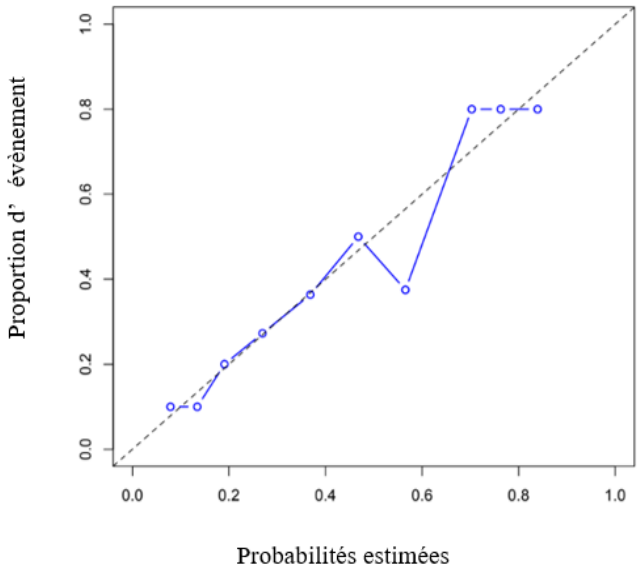



Figure 1: Exemple de diagramme de fiabilité. Un point se trouvant au-dessous de la bissectrice représente une probabilité surestimée et un point se trouvant au-dessus de la bissectrice représente une probabilité sous-estimée.

L'Expected Calibration Error (ECE- Naeini & al. (2015)) permet d'évaluer globalement la calibration *in the small* par une valeur numérique. Il est calculé comme la moyenne pondérée des valeurs absolues des écarts des taux observés à la moyenne des probabilités estimées dans chaque groupe :

$$ECE = \sum_{g=1}^{G} \frac{n_g}{n} |\overline{y}_g - \overline{p}_g|$$

Une bonne calibration correspond à un ECE faible, soit proche de 0.

Une mesure qui permet d'évaluer à la fois la discrimination et la calibration est le score de Brier. C'est la moyenne des carrés des différences entre les probabilités estimées et les

valeurs de $y$ pour chaque individu (Brier, G.W. (1950)) :

$$B = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2$$

Le score de Brier varie entre $[0; 1]$. Un score plus faible représente une meilleure performance.

Plusieurs décompositions du score de Brier existent en pratique dont l'une des premières en 2 éléments est donnée par Sanders, F. (1963) :

$$B = \frac{1}{n}\sum_{g=1}^{G} n_g \overline{y}_g(1 - \overline{y}_g) + \frac{1}{n}\sum_{g=1}^{G} n_g(p_g - \overline{y})^2 \tag{1}$$

Le premier terme de l'équation (1) est appelé *résolution* et le second terme correspond à la *calibration*. Vu que la résolution est liée à la discrimination (Huang, C. & al. (2021)), le score de Brier est généralement considéré comme une mesure de la qualité globale d'un classifieur.

Ainsi, nous utilisons l'AUC pour évaluer la discrimination ; la calibration en moyenne, le diagramme de fiabilité et l'ECE pour évaluer la calibration et le score de Brier pour les deux à la fois.

# 3 Méthodologie de simulation

Notre simulation vise à évaluer la sensibilité des indices cités dans la section précédente dans des conditions où la calibration est de plus en plus mauvaise. L'objectif est de comprendre comment ces indices réagissent aux déviations des probabilités estimées par rapport aux vraies probabilités, ce qui peut être crucial dans des applications réelles où la calibration des prédictions est nécessaire.

Pour ce faire, nous avons généré quatre vecteurs $x_j$ ($j \in \{u, asym, unif, cloche\}$) de taille $n = 2000$ à partir de différentes lois (beta, normale et logistique). Chacun de ces vecteurs $x_j$ est ensuite transformé en un vecteur $\pi_j$[1] de probabilités en utilisant le modèle logistique $\pi_j = \frac{e^{\beta x_j}}{e^{\beta x_j}+1}$ avec $\beta = 1$. Ces 4 vecteurs de probabilités présentent des distributions avec des formes différentes, comme illustrées dans la figure 2 :

- Forme en $u$ : symétrique avec 2 modes aux extrêmes.
- Forme *asymétrique* : étalée à gauche (asymétrique positive).
- Forme *uniforme* : toutes les valeurs ont la même probabilité d'occurrence.
- Forme en *cloche* : symétrique avec mode à 0.5.

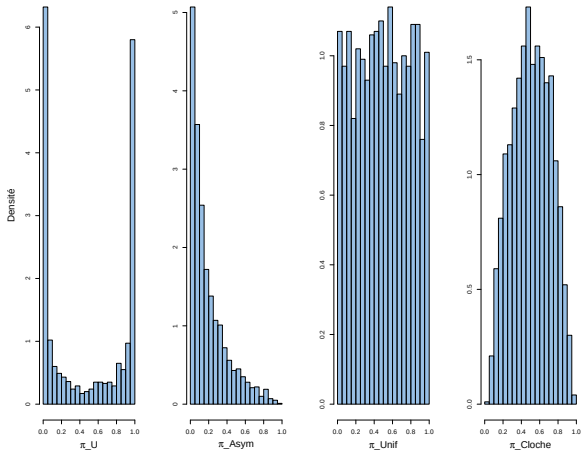



Figure 2: Distributions des probabilités $\pi$. De droite à gauche : *u*, *asymétrique* positive, *uniforme* et *cloche*

Pour la suite, 1000 valeurs binaires $y_{ijk}$ ($k \in \{1, \ldots, 1000\}$), issues de lois de Bernoulli de paramètre $\pi_{ij}$, sont générées de façon aléatoire.

Afin d'obtenir des vecteurs de probabilités estimées $p_j$ que nous considérons issus de classifieurs probabilistes, nous dévions les 4 vecteurs de vraies probabilités $\pi_j$ en utilisant 3 méthodes de déviation qui présentent des caractéristiques différentes (figure 3). Pour chaque méthode, on varie l'intensité de l'erreur :

- **Méthode 1** : la valeur du coefficient $\beta$ est remplacée par les valeurs 0.75, 0.5 et 0.25. Cette méthode donne des déviations monotones, symétriques, plus accentuées sur les valeurs extrêmes.
- **Méthode 2** : des bruits gaussiens de moyenne nulle et d'écart-type 0.2, 0.5, 1 et 2 sont ajoutés aux vecteurs $\pi_j$. Cette méthode donne des déviations non monotones, symétriques, plus accentuées sur les valeurs centrales.
- **Méthode 3** : le premier tercile de la distribution des $\pi_j$ est multiplié par 1.075, 1.15, 1.3, 1.4, 1.5 et le dernier tercile par 0.925, 0.85, 0.7, 0.6, 0.5 (inspiré de Huang, Y. & al. (2020)). Cette méthode donne des déviations non monotones, asymétriques, plus accentuées sur les valeurs proches de 1.

Enfin, nous calculons les indices de performance, à savoir l'AUC, le score de Brier, l'ECE et la calibration en moyenne. Pour ce faire, nous comparons, pour chaque répétition $k$, les valeurs $y_j$ de la réponse à une situation de référence (valeur optimale) où le modèle estime parfaitement les vraies probabilités (comparaison des valeurs $y_j$ de la réponse aux vraies probabilités $\pi_j$).

[1]L'élément générique $i$ du vecteur $\pi_j$ est considéré comme la vraie probabilité associée à l'$i$ème observation.

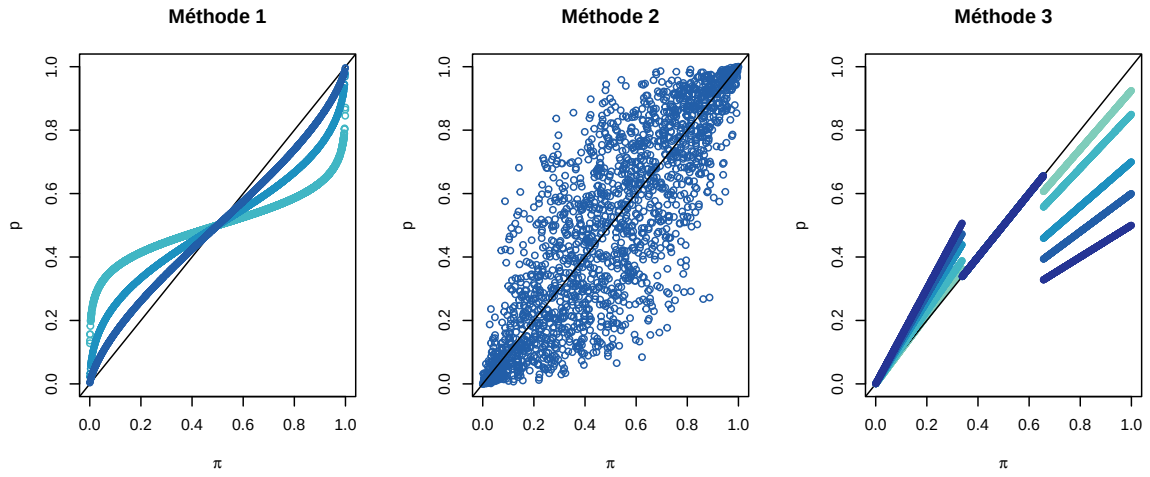



Figure 3: Caractéristiques des écarts des probabilités estimées $p$ aux vraies probabilités $\pi$. Pour la méthode 2, $sd = 1$

Les valeurs $y_j$, pour chaque répétition $k$, sont ensuite comparées aux probabilités estimées $p_j$ issues du croisement des 4 distributions des vraies probabilités et des déviations de tout type et intensité.

# 4 Résultats et discussion

La figure 4 nous donne une visualisation de l'influence de la distribution des probabilités sur les valeurs optimales de l'AUC et du score de Brier, en l'absence de toute déviation. Ces valeurs sont étroitement liées à la distribution des probabilités. Lorsque la distribution présente plus de valeurs extrêmes (forme en $u$), la valeur optimale tend à être meilleure. En revanche, lorsque la distribution est plus concentrée autour des valeurs centrales (forme en *cloche*), la valeur optimale tend à être moins bonne.

Les figures 5 et 6 illustrent respectivement la sensibilité de la calibration en moyenne et de l'ECE à différents niveaux d'écarts entre les probabilités estimées et les vraies probabilités, en fonction des distributions des probabilités et des méthodes de déviation.

La mesure de la calibration *in the large* est plus sensible lorsque la distribution de probabilités est *asymétrique* ou l'ampleur des écarts est asymétrique (figure 5). Plus spécifiquement, pour la méthode 3 présentant des déviations asymétriques, la calibration en moyenne est sensible pour toutes les formes de distribution, avec une sensibilité plus importante pour la distribution *asymétrique*. Ce qui n'a pas été noté avec les méthodes 1 et 2 (qui génèrent des écarts symétriques) où l'indice est sensible seulement lorsque la forme de distribution est *asymétrique*.

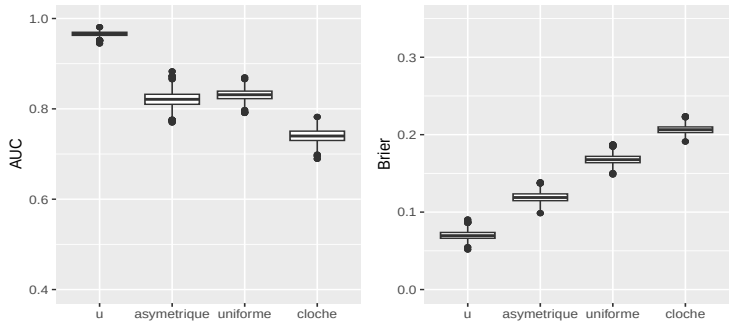



Figure 4: Boîtes à moustaches des 1000 valeurs optimales de l'AUC et du score de Brier pour chaque forme de distribution

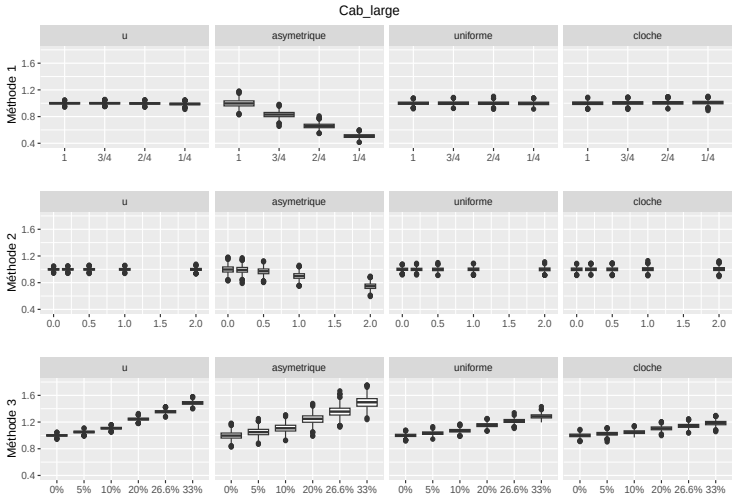



Figure 5: Boîtes à moustaches des 1000 valeurs obtenues de la calibration en moyenne par méthode et forme de distribution

L'ECE se révèle plus sensible lorsque les écarts [2] les plus importants se situent au niveau des probabilités présentant une densité plus élevée (figure 6). Pour la méthode 1, les plus larges écarts se trouvent au niveau des valeurs extrêmes (figure 3), ce qui rend l'ECE plus sensible pour les distributions en *u* et *asymétrique*. Pour la méthode 2, les plus grandes déviations se trouvent au niveau des valeurs centrales, rendant l'ECE plus sensible pour les distributions des vraies probabilités en *cloche* et *uniforme*. Enfin, avec la méthode 3, les plus larges écarts sont au niveau des probabilités proches de 1, ce qui fait qu'on observe une plus grande sensibilité pour la distribution des vraies probabilités en *u* par rapport au reste. Cela est d'ailleurs plus marquant en regardant la distribution des vraies probabilités *asymétrique* présentant une très faible sensibilité étant donné qu'elle est concentrée au niveau

[2] écarts entre les probabilités estimées et les vraies probabilités

des probabilités proches de 0.

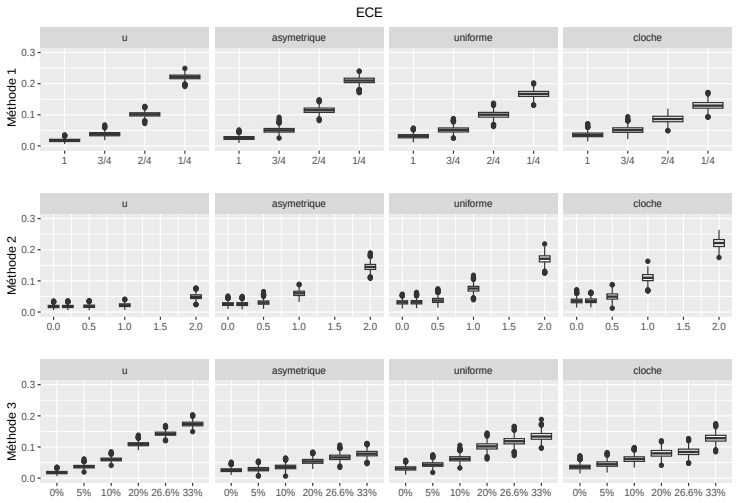



Figure 6: Boîtes à moustaches des 1000 valeurs obtenues de l'ECE par méthode et forme de distribution

Pour vérifier si une étude plus approfondie de la calibration pouvait nous aider à mieux détecter les estimations des probabilités de différente intensité par rapport à un indice global comme l'ECE, nous avons construit les diagrammes de fiabilité correspondants (voir figure 7).

Les diagrammes de fiabilité confirment que l'interprétation de la qualité de classifieurs peut changer selon la forme de la distribution des vraies probabilités lorsque la qualité des estimations reste la même.

Par exemple, à parité de type (méthode 2) et d'intensité ($sd = 1$ : assez importante) de la déviation, le diagramme peut générer des interprétations très différentes de la qualité de la calibration selon la forme de la distribution des vraies probabilités. Si la forme est en $u$, l'estimation des probabilité semble presque parfaite, tandis que pour les autres formes la calibration ne semble pas satisfaisante.

# 5 Conclusion et perspectives

Cette étude sur la sensibilité des indices de qualité d'un classifieur probabiliste a permis de réaliser que ces indices dépendent non seulement de l'intensité des écarts des probabilités estimées aux vraies probabilités, mais aussi des caractéristiques de ces écarts et des vraies probabilités. Même si en théorie l'AUC et le score de Brier sont compris respectivement dans les intervalles $[0.5; 1]$ et $[0; 1]$, dans la pratique les valeurs optimales de ces indices dépendent de la distribution des vraies probabilités $\pi$, qui est inconnue en pratique. De plus, la vraie qualité de la calibration peut ne pas être détectée avec les indices considérés dans ce papier vu qu'ils dépendent de la distribution de $\pi$.

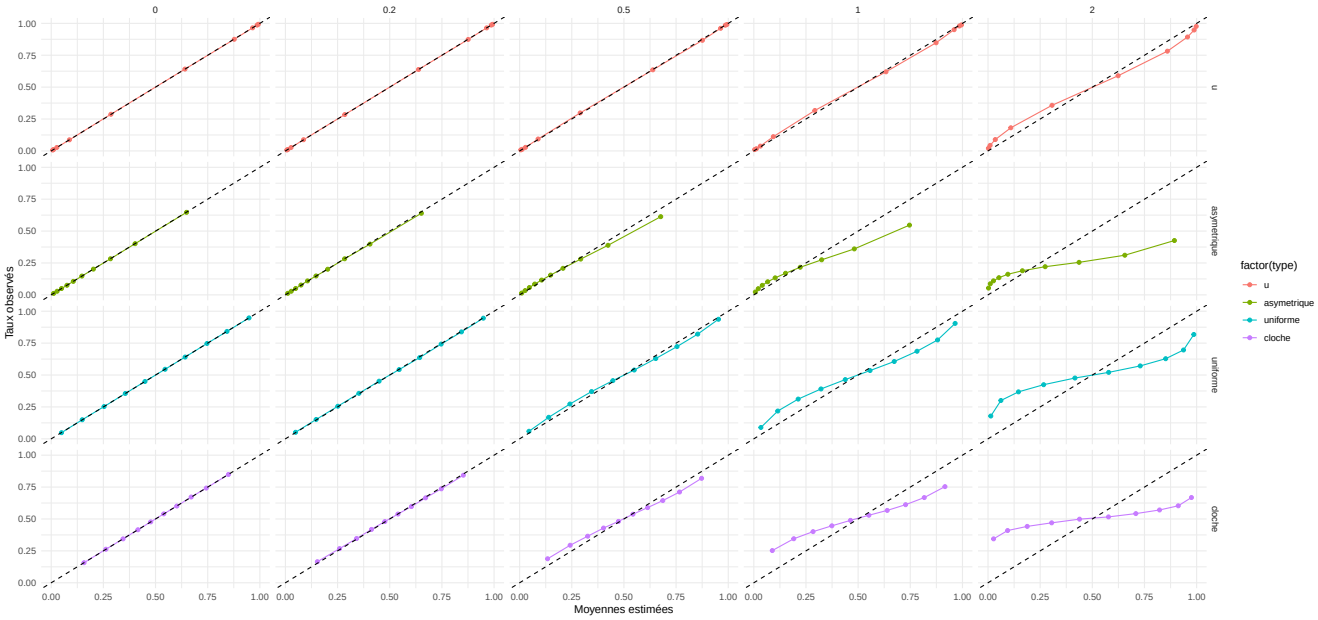



Figure 7: Diagramme de fiabilité pour chaque forme de la distribution des vraies probabilités et chaque niveau des déviations issues de la méthode 2

Il convient donc d'être prudent dans l'interprétation de la qualité des classifieurs probabilistes avec ces indices afin d'éviter des interprétations erronées ou biaisées. Des recherches futures sont nécessaires pour développer des méthodologies plus robustes permettant d'évaluer de manière fiable les performances des classifieurs probabilistes.

# Bibliographie

Brier, G.W. (1950), *Verification of forecasts expressed in terms of probability*, Monthly Weather Review, 78 (1): 1-3.

Calders, T. and Jaroszewicz, S. (2007), *Efficient AUC Optimization for Classification*, Knowledge Discovery in Databases: PKDD 2007, Lecture Notes in Computer Science, vol 4702, Springer, Berlin, Heidelberg

Huang, Y. and Li, W. and Macheret, F. and Gabriel, R. A and Ohno-Machado, L. (2020), *A tutorial on calibration measurements and calibration models for clinical prediction models*, JAMIA, 27(4), pp. 621-633

Huang, C. and Li, S.X. and Caraballo, C. and Masoudi, F.A. and Rumsfeld, J.S. and Spertus, J.A. and Normand, S.T., Mortazavi, B.J. and Krumholz, H.M. (2021), *Performance Metrics for the Comparative Analysis of Clinical Risk Prediction Models Employing Machine Learning*, Circ Cardiovasc Qual Outcomes, 14(10):e007526

Pepe, M. and Janes, H. (2013), *Methods for Evaluating Prediction Performane of Biomarkers*

*and Tests*, In: Lee, ML., Gail, M., Pfeiffer, R., Satten, G., Cai, T., Gandy, A. (eds) Risk Assessment and Evaluation of Predictions. Lecture Notes in Statistics, vol 215. Springer, New York, NY

Naeini, P.M. and Cooper, G. and Hauskrecht, M. (2015), *Obtaining Well-Calibrated Probabilities Using Bayesian Binning*. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1)

Sanders, F. (1963), *On Subjective Probability Forecasting*, J. Appl. Meteor. Climatol., 2

Van Calster, B. and Daan, N. and Yvonne, V. and Bavo, C. and Michael, J.P. and Ewout, W. S. (2016), *A calibration hierarchy for risk models was defined: from utopia to empirical data*, Journal of clinical epidemiology

# Données omiques

# L'IMPACT NÉGATIF DES MATRICES DE RÉFÉRENCE INCOMPLÈTES SUR LA PERFORMANCE DE LA DÉCONVOLUTION DES FRÉQUENCES CELLULAIRES À PARTIR DE L'EXPRESSION GÉNIQUE

Kalidou BA [1,2], Rodolphe THIÉBAUT [1,2,3], Xavier HINAUT[4,5,6] & Boris HEJBLUM [1,2]

[1] *Univ. Bordeaux, INSERM, INRIA, SISTM team, BPH, U1219, F-33000 Bordeaux, France*

[2] *Vaccine Research Institute, F-94000 Créteil, France*

[3] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*

[4] *INRIA Bordeaux Sud-Ouest, France*

[5] *LaBRI, Bordeaux INP, CNRS, UMR 5800*

[6] *Institut des Maladies Neurodégénératives, Université de Bordeaux, CNRS, UMR 5293*

**Résumé.** La déconvolution cellulaire désigne l'estimation des fréquences des populations cellulaires à partir des mesures de l'expression des gènes dans un échantillon biologique. Bien que de nombreuses approches supervisées aient été proposées pour résoudre ce problème (telles que `CibersortX` – Newman et al. (2019), ou `scaden` – Menden et al. (2020)), leurs bonnes performances dépendent essentiellement de la matrice des signatures d'expression génique de référence par population. Cette matrice encode les profils d'expression génique de référence des différents types cellulaires, à partir de connaissances préalables et de jeux de données externes. Toutefois, le cas où certaines populations cellulaires présentes dans l'échantillon sont manquantes dans la matrice de référence n'a reçu qu'une attention limitée dans la profusion d'algorithmes de déconvolution proposés, en particulier au vu de sa réalité pratique. Nous quantifions le manque de robustesse des méthodes de déconvolution de l'état de l'art, à la fois dans des simulations numériques et aussi à l'aide de jeux de données réelles. Nos simulations se basent sur une distribution multivariée (soit de Poisson soit Gaussienne) au plus proche de données réelles extraites de la littérature. Nos résultats démontrent que les performances de déconvolution restent relativement inchangées tant que la matrice de référence inclut la grande majorité des populations cellulaires présentes dans l'échantillon, mais qu'à l'inverse les performances de toutes les méthodes de déconvolution se détériorent rapidement à mesure que la matrice de référence devient de plus en plus incomplète. Cependant l'impact des populations cellulaires manquantes dans la matrice de référence dépend de leur fréquence réelle dans l'échantillon.

**Mots-clés.** Déconvolution cellulaire, RNA-Seq, Matrices de signatures de référence, Types cellulaires, Décomposition matricielle, Optimisation.

**Abstract.** Cellular deconvolution refers to the estimation of cellular population frequencies from gene expression measurements in a biological sample. While numerous supervised approaches have been proposed (such as `CibersortX` or `scaden`), their good performance critically depends on the reference signature matrix. This matrix encodes the gene expression profiles of the different cell types from external prior knowledge. However, addressing the common scenario of missing cellular populations from the reference matrix has received limited attention compared to the profusion of proposed deconvolution algorithms. We assess the lack of robustness of the state-of-the-art deconvolution methods in both simulations and benchmarking real data. Our simulations designs, based

on either a Poisson or a Gaussian multivariate distribution, are validated against real data from the literature. Results from simulations and multiple real datasets, demonstrate that deconvolution performance remains relatively unaffected as long as the reference matrix includes most cellular populations present in the sample. Conversely, performance rapidly deteriorates for all deconvolution methods as the reference matrix becomes increasingly incomplete. Moreover, the impact of missing cell populations in the reference signature matrix depends on their actual frequency in the sample.

**Keywords.** Cellular deconvolution, RNA-seq, Reference signature matrix, Cell type, Matrix decomposition, Optimisation.

# 1 Introduction

La déconvolution cellulaire désigne l'estimation des abondances cellulaires à partir de profils d'expression génique spécifiques à chaque type cellulaire grâce à des données de séquençage de l'ARN en masse (*bulk RNA-seq data*) (Avila et al., 2020). L'abondance et la fréquence des populations cellulaires dans un tissu peuvent varier selon différents facteurs, tels que le stade de développement, l'état pathologique ou encore l'influence de l'environnement. La déconvolution des signaux d'expression génique mélangés permet alors de déterminer la contribution spécifique de chaque type de cellule à l'expression globale des gènes. C'est un problème difficile, en raison de la complexité et de l'hétérogénéité des tissus biologiques, de la grande dimension des données d'expression génique, de l'incertitude sur les signatures de référence d'expression génique ainsi que sur les types cellulaires pertinents à inclure, et de la présence de bruit technique dans les mesures RNA-seq (Wang et al., 2019).

Les approches supervisées s'appuient principalement sur les signatures de référence disponibles pour faire correspondre les profils d'expression génique observés selon un modèle de mélange linéaire. Il est impératif de connaître précisément la matrice des signatures d'expression de référence. En effet, celle-ci joue un rôle majeur dans la déconvolution. Cette matrice est essentiellement une représentation des profils d'expression génique pour des types de cellules connus (Shen-Orr and Gaujoux, 2013). La précision, la pertinence, l'adéquation et l'exhaustivité de la matrice de référence par rapport aux données étudiées sont essentielles pour obtenir de bons résultats par ces méthodes de déconvolution. Un problème important et fréquent dans la déconvolution cellulaire est celui des populations manquantes dans la matrice de signature de référence. En effet, en l'absence d'une référence correspondante pour un type de cellule particulier présent dans l'échantillon biologique, le modèle d'estimation se retrouve mal spécifié et les résultats sont alors biaisés (plus ou moins largement). La normalisation induite par les modèles de mélange implique que l'absence de populations cellulaires dans la matrice de référence affecte aussi l'estimation des autres types de cellules connus. Cette limitation rend difficile l'utilisation de ces méthodes de déconvolution dans des contextes où les populations cellulaires présentes sont inconnues (par exemple dans des échantillons sanguins), et empêche l'identification de populations cellulaires rares ou nouvelles avec ces méthodes.

Afin de mieux comprendre les implications d'une matrice de référence de signature incomplète sur la performance des méthodes de déconvolution, nous examinons un ensemble de 16 méthodes extraites d'une liste exhaustive d'une trentaine de méthodes sélectionnées sur la base des éléments suivants : i) la possibilité de personnaliser *a priori* les informations contenues dans la matrice de référence ; ii) l'absence de restrictions sur le nombre de types

cellulaires identifiés et la robustesse de la détection des populations de faible abondance ; et iii) l'intégration de la contrainte de non-négativité et la standardisation des abondances estimées. La phase de sélection préliminaire a été orientée par le nombre de citations sur PubMed, en tenant compte des dates de publication. Nous avons veillé à représenter le large éventail de méthodologies proposées pour traiter la déconvolution, y compris des méthodes linéaires, quadratiques, probabilistes et d'apprentissage automatique. Par la suite, nous nous sommes concentrés exclusivement sur les méthodes démontrant les meilleures performances dans l'estimation des populations cellulaires largement étudiées dans la littérature et pertinentes dans l'étude du système immunitaire à partir du sang circulant (telles que les cellules B, les cellules T CD4+, les cellules T CD8+, les neutrophiles et les plasmocytes).

# 2 Méthodes

## 2.1 Méthodes de déconvolution

Nous avons évalué plusieurs méthodes de déconvolution cellulaire basées sur des matrices de signature de référence prédéfinies, chacune employant des approches et des techniques spécifiques pour la déconvolution des types de cellules. Dans cet ensemble, `CibersortX` (Steen et al., 2020), `scaden` (Menden et al., 2020) et `Fardeep` (Hao et al., 2019) se distinguent par leur capacité à estimer avec précision les proportions de types cellulaires de manière robuste, grâce à l'utilisation de techniques d'apprentissage automatique. Plus précisément, ils utilisent des techniques d'apprentissage automatique telles que la régression vectorielle de support, les réseaux neuronaux d'apprentissage profond et un algorithme des moindres carrés *trimmés* adaptatifs. En particulier, les méthodes de déconvolution basées sur l'apprentissage profond telles que `scaden` intègrent une phase de simulation de données artificielles pour optimiser le réseau, et `Fardeep` est robuste en présence de bruit en supprimant les valeurs aberrantes avant la déconvolution en utilisant l'idée des moindres carrés trimmés. `AutoGeneS` (Aliee and Theis, 2021) et `DeconRNASeq` (Gong and Szustakowski, 2013) imposent des contraintes de non-négativité et intègrent respectivement les approches des moindres carrés et de la programmation quadratique. `LinDeconSeq` (Li et al., 2020) utilise des méthodes de notation de la spécificité et de linéarité mutuelle pour la sélection des gènes marqueurs, puis applique une régression linéaire robuste pondérée. `EPIC` (Racle and Gfeller, 2020), spécialement conçue pour la déconvolution avec des populations cellulaires inconnues, effectue une régression par les moindres carrés. Les cadres bayésiens ont été adoptés par `BayesPrism` (Chu et al., 2022) et `CDSeq` (Kang et al., 2019), ce dernier mettant en œuvre un modèle bayésien hiérarchique. `DCQ` (Zeev et al., 2014) et `ABIS` (Monaco et al., 2019) utilisent respectivement la régularisation *elastic-net* et le modèle linéaire robuste. QProg (sans contraintes) et QProgwc (Gong et al., 2011) (avec contrainte de non-négativité et de somme à 1) utilisent la régression quantile, prenant en compte l'hétérogénéité spécifique au type de cellule dans les données d'expression génique. `LR` applique la régression linéaire régularisée, tandis que `RLS` (Sharma et al., 2019) se concentre sur la sélection de régions de données d'expression génique qui présentent des modèles similaires entre les échantillons.

## 2.2 Simulation des données

Nous avons effectué des simulations numériques pour étudier les performances des méthodes présentées ci-dessus. Un grand nombre de populations cellulaires a été considéré dans la matrice de référence $W$. Pour couvrir un large panel, nous avons simulé nos données selon deux distributions : gaussienne d'une part (dont les paramètres sont obtenus par une approximation gaussienne de la matrice de signature *LM22* proposée par (Chen et al., 2018)) et Poisson d'autre part (qui permet de simuler des données de comptage à l'instar des données RNA-Seq réelles). Afin de générer de la corrélation, inhérente entre les proportions des populations cellulaires (de par leur nature compositionnelle), nous avons défini une matrice de variance-covariance uniforme pour la distribution gaussienne, et adopté la stratégie proposée par (Barbiero and Ferrari, 2015) pour la distribution de Poisson. Les proportions de types de cellules de tous les individus en vrac $P$ sont générées avec une distribution de Dirichlet. Enfin, un bruit indépendant $\epsilon$ est ajouté pour générer les données d'expression génique $X(X = WP+\epsilon)$. La Table 1 résume ces choix. La Figure 1A présente un exemple de ces simulations, comparé aux données réelles.

| | |
|---|---|
| Nombre de gènes | $n = 500$ |
| Nombre de population de cellules | $K = 100$ |
| Number d'échantillon | $N = 50$ |
| Matrice des proportions | $P_{i,*} \sim \mathcal{DP}(K, \alpha_k)$, où $\alpha_k \in [1, ..., K]$, $\sum_{j=1}^{K} P_{i,j} = 1$ et $\forall i \in [1, ..., N]$ $P_{i,j} < 1,\ \forall\ i, j$ |
| Avec une distribution Gaussienne | $W \sim e^{\mathcal{N}(\mu,\, \Sigma)} + 6$, avec $\mu \sim \mathcal{U}(-1,\, 0.5)$<br>$\epsilon \sim \mathcal{N}(0, \sigma)$ où $\sigma \sim \mathcal{U}(1, 2)$ |
| Avec une distribution de Poisson | $W \sim \mathcal{P}(\lambda_k, \Sigma)$, avec $\lambda_k \sim \mathcal{U}(10, 25)$<br>$\epsilon \sim \mathcal{P}(\theta_k)$ où $\theta_k \sim \mathcal{U}(0, 0.3)$ |

TABLE 1 : Configuration des simulations numériques

Trois scénarios ont ensuite été définis pour étudier les performances des différentes méthodes de déconvolution en fonction du nombre de populations cellulaires dans la matrice de référence $W$. Le scénario « *randomly* » consiste à extraire aléatoirement (sans remise) un pourcentage de populations cellulaires définient dans la matrice de référence $W$. Les scénarios « *lowest* » et « *highest* » correspondent à l'extraction à partir de la matrice de référence les populations de cellules ayant une abondance moyenne soit faible, soit élevée, respectivement.

## 2.3 Métriques d'évaluation

La précision des différentes méthodes de déconvolution est évaluée à l'aide des mesures de l'erreur quadratique moyenne relative ($RRMSE$) et du coefficient de corrélation intra-classe ($ICC$) entre les proportions estimées et les véritables proportions.

Le $RRMSE$ fournit une mesure normalisée et interprétable de l'erreur de prédiction, qui facilite des comparaisons équitables entre différents modèles, ensembles de données et échelles, même lorsque l'abondance réelle est faible.

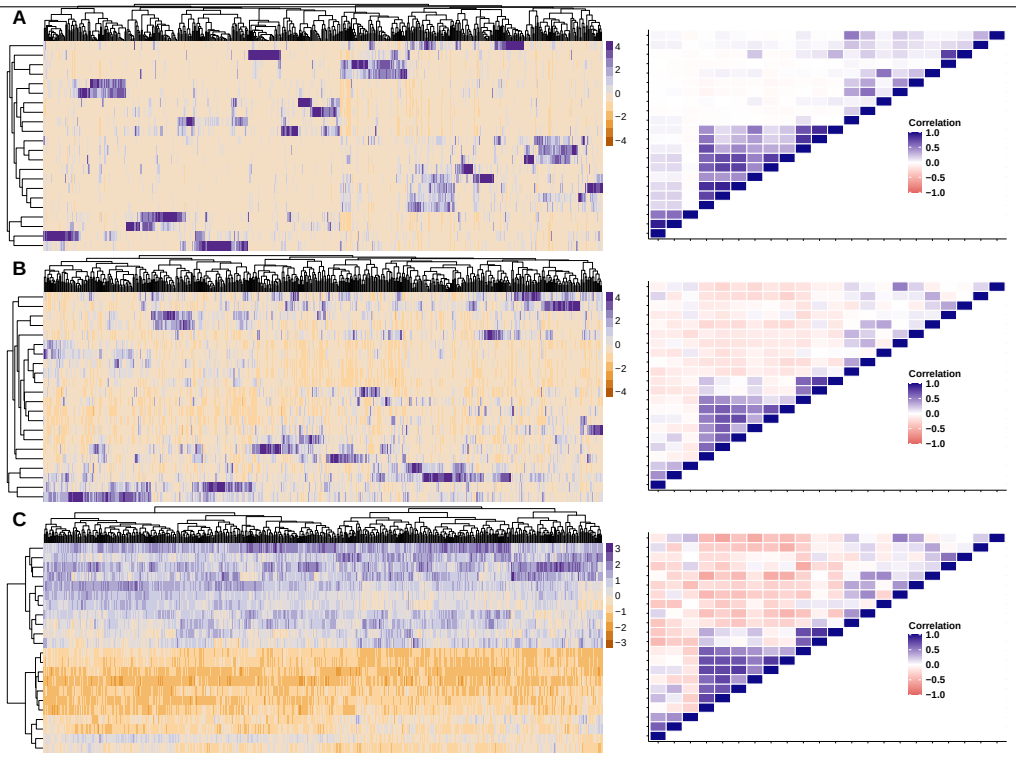



FIGURE 1 : Analyse comparative de matrices de référence simulées issues de distributions gaussiennes **(B)** et de Poisson **(C)**, ainsi que de la matrice de signature de référence *LM22* **(A)**. Les lignes représentent les populations de cellules (K = 22) et les colonnes représentent les gènes (N = 547). À droite, les cartes thermiques illustrent les corrélations par paire entre les populations de cellules, les valeurs R-carré étant utilisées comme mesure de l'association. Les corrélations les plus fortes sont représentées en bleu, tandis que les corrélations les plus faibles sont représentées en rouge.

L'$ICC$ (compris entre 0 et 1) est une alternative au coefficient de corrélation de Pearson (ce dernier étant sensible à la présence de valeurs aberrantes) et mesure la fiabilité et la cohérence des mesures, largement utilisé pour les interprétations cliniques. Plus il est proche de 1, meilleure est la performance. Nous utilisons la définition de l'$ICC_3$ proposée par (PE and JL, 1979), car nous sommes intéressés par l'évaluation d'un ensemble fixe de treize méthodes de déconvolution. Ces deux indicateurs se calculent ainsi pour une population spécifique :

$$RRMSE = \frac{\sqrt{\frac{\sum_{j=1}^{J}(\hat{f}_j - f_j)^2}{J}}}{\sigma_f} \qquad ICC = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2} \tag{1}$$

où $j$ est l'indice de l'échantillon et $J$ le nombre total d'échantillons biologique dans un jeu de données. $\hat{f}_j$ est la proportion estimée de la population de cellules de l'échantillon $j$ et $f_j$ est la vérité terrain de l'échantillon $j$. $\sigma_r^2$, $\sigma_c^2$ et $\sigma_e^2$ correspondent respectivement aux variances de l'écart par rapport à la moyenne pour l'échantillon $j$, au biais de la mesure $k$ et à la composante résiduelle.

# 3 Résultats

En considérant toutes les populations cellulaires lors de la déconvolution, nous avons observé que 5 des 16 méthodes, `scaden`, `EPIC`, `DCQ`, `BayesPrism`, et `RLS`, avaient des $ICC$ moyen inférieurs à $0,5$ ; contrairement aux autres méthodes qui fournissaient des estimations bien meilleures. Ces résultats pourraient s'expliquer par le fait que des méthodes telles que `scaden` n'ont pas été optimisées pour traiter plus d'une vingtaine de populations cellulaires.

Indépendamment de la stratégie de simulation adoptée et de la méthode de déconvolution utilisée, plus le nombre de populations cellulaires identifiées dans la matrice de signature est faible, plus les $ICC$ observées entre les véritable valeurs et les prédictions de la déconvolution sont faibles. Cette dégradation des performances était plus importante lorsque les populations cellulaires manquantes étaient très fréquentes dans le mélange, comme le montrent les résultats obtenus avec les 50 populations cellulaires les plus abondantes. La performance moyenne de l'ICC était supérieure à $0,65$ pour le scénario "lowest", alors qu'elle était estimée supérieure à $0,5$ avec les 50 populations cellulaires les moins abondantes dans le scénario "highest". Dans le scénario "lowest", un grand nombre ($> 20\%$) de population doit être considéré comme inconnu pour avoir un impact significatif sur la performance. Ces constatations se reflètent également dans le $RRMSE$ qui augmente à chaque niveau d'incomplétude de la matrice de référence. Ces résultats sont détaillés dans les Figures 2 et 3.

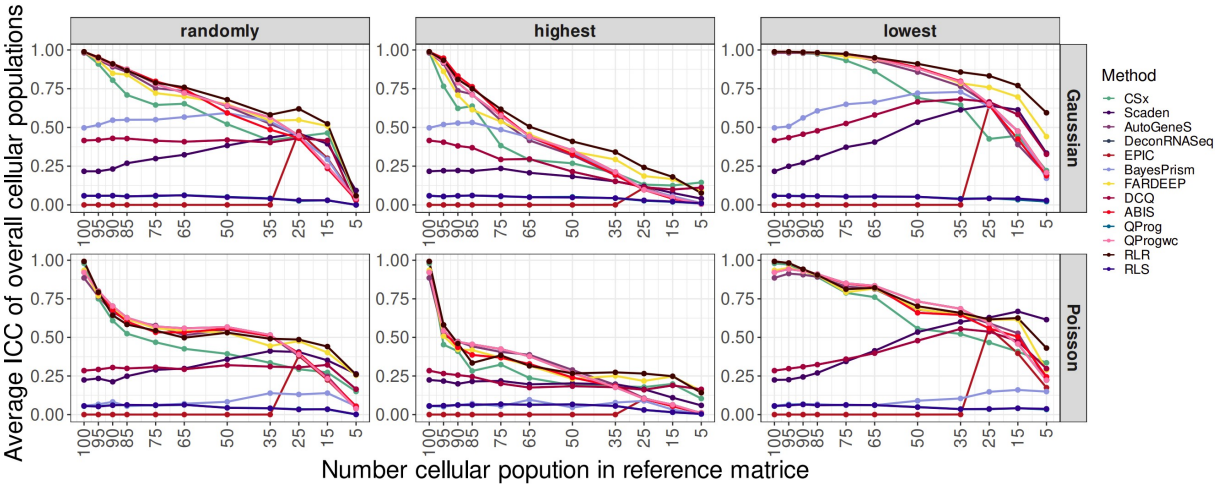



FIGURE 2 : Coefficient de corrélation intraclasse (ICC) moyen des populations cellulaires globales

# 4 Discussion

Nos résultats de simulation ont permis de mieux comprendre l'impact d'une matrice de référence incomplète. Nous avons constaté que les performances des méthodes de déconvolution se détérioraient à mesure que le nombre de populations cellulaires manquantes

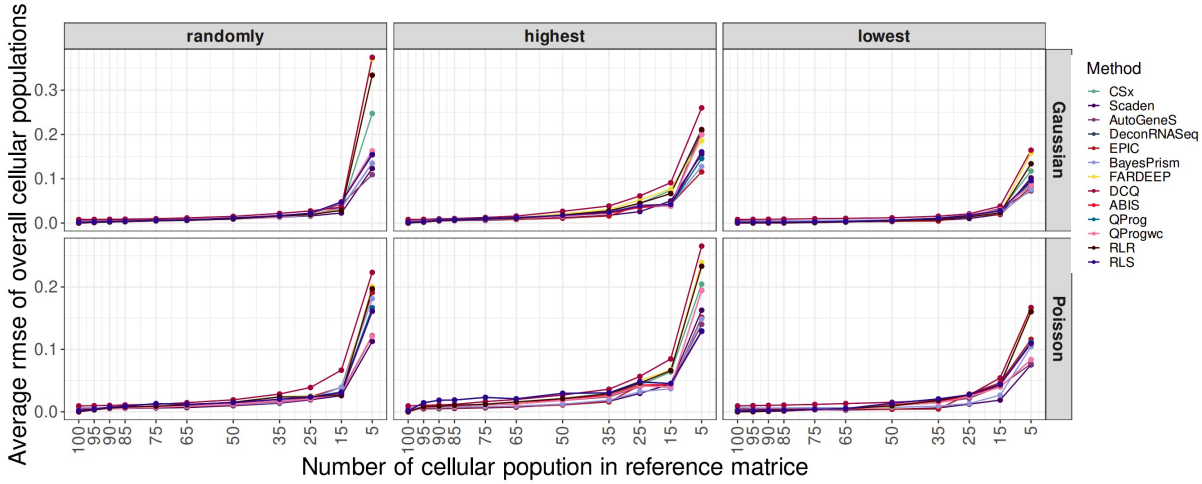



FIGURE 3 : Erreur quadratique moyenne relative (RRMSE) des populations cellulaires globales.

dans la matrice de référence augmentait, d'autant plus que les populations cellulaires manquantes étaient fréquentes dans le mélange. Cela souligne la sensibilité des méthodes de déconvolution à la disponibilité de matrices de signatures de référence complètes.

Il convient de noter que certaines méthodes de déconvolution ont montré des limites lorsqu'elles étaient confrontées à un grand nombre de populations cellulaires, comme nous l'avons observé dans nos simulations. Des méthodes comme `scaden`, `EPIC`, `DCQ`, `BayesPrism` et `RLS` ont eu du mal à fournir des estimations précises lorsque la matrice de référence contenait un grand nombre de populations cellulaires.

En conclusion, notre étude met en évidence la forte dépendance entre la performance des méthodes de déconvolution cellulaire et la qualité et l'exhaustivité des matrices de signature de référence. Les résultats soulignent la nécessité de développer des algorithmes de déconvolution plus robustes, capables de traiter des tissus divers et complexes. Les recherches futures devraient se concentrer sur l'amélioration de l'adaptabilité et de la précision des méthodes de déconvolution, en particulier en présence de matrices de référence incomplètes. En outre, les efforts visant à normaliser les matrices de référence et à établir les meilleures pratiques pour la construction des matrices de référence peuvent améliorer la reproductibilité et la comparabilité des études de déconvolution dans différents contextes biologiques. En résumé, la déconvolution cellulaire est un outil puissant pour disséquer les compositions tissulaires complexes, mais son efficacité dépend de la qualité des données de référence et des capacités de la méthode de déconvolution choisie.

# Références

H Aliee and FJ Theis. Autogenes : Automatic gene selection using multi-objective optimization for rna-seq deconvolution. *Cell Systems*, 12(7):706–715, 2021.

CF Avila, J Alquicira-Hernandez, JE Powell, P Mestdagh, and PK De. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11 (1):1–4, 2020.

A Barbiero and PA Ferrari. Simulation of correlated poisson variables. *Applied Stochastic Models in Business and Industry*, 31:669–680, 2015.

B Chen, MS Khodadoust, CL Liu, AM Newman, and AA Alizadeh. Profiling tumor infiltrating immune cells with cibersort. *Methods in Molecular Biology*, 1711:243–259, 2018.

T Chu, Z Wang, D Pe'er, and CG Danko. Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology. *Nature Cancer*, 3(4):505–517, 2022.

T Gong and JD Szustakowski. Deconrnaseq : a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–5, 2013.

T Gong, N Hartmann, I S Kohane, V Brinkmann, F Staedtler, M Letzkus, S Bongiovanni, and J D Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLOS One*, 6(11):e27156, 2011.

Y Hao, M Yan, BR Heath, YL Lei, and Y Xie. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLOS Computational Biology*, 15(5):e1006976, 2019.

Q Kang, K andMeng, I Shats, DM Umbach, M Li, Y Li, X Li, and L Li. Cdseq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Computational Biology*, 15(12):e1007510, 2019.

H Li, A Sharma, W Ming, X Sun, and H Liu. A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. *BMC Genomics*, 21:652, 2020.

K Menden, M Marouf, S Oller, A Dalmia, DS Magruder, K Kloiber, P Heutink, and S Bonn. Deep learning-based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, 2020.

G Monaco, B Lee, W Xu, S Mustafah, YY Hwang, C Carré, N Burdin, L Visan, M Ceccarelli, M Poidinger, A Zippelius, J Pedro de Magalhães, and A Larbi. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell Reports*, 26(6):1627–1640, 2019.

AM Newman, CB Steen, CL Liu, AJ Gentles, AA Chaudhuri, F Scherer, MS Khodadoust, MS Esfahani, BA Luca, D Steiner, and M Diehn. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

Shrout PE and Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86:420–8, 1979.

J Racle and D Gfeller. Epic: A tool to estimate the proportions of different cell types from bulk gene expression data. *Methods in molecular biology*, 15:233–248, 2020.

A Sharma, E Merritt, X Hu, A Cruz, C Jiang, H Sarkodie, Z Zhou, J Malhotra, GM Riedlinger, and S De. Non-genetic intra-tumor heterogeneity is a major predictor of phenotypic heterogeneity and ongoing evolutionary dynamics in lung tumors. *Cell Reports*, 29(8):2164–2174, 2019.

SS Shen-Orr and R Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, 25(5):571–8, 2013.

CB Steen, CL Liu, AA Alizadeh, and AM Newman. Profiling cell type abundance and expression in bulk tissues with cibersortx. *Methods in Molecular Biology*, pages 135–157, 2020.

T Wang, B Li, CE Nelson, and S Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC Bioinformatics*, 20(1):40, 2019.

A Zeev, S Yael, D Eyal, Zohar BI, V Liran, KS Hadas, M Tal, M Ella, M Michal, GV Irit, and A Ido. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular Systems Biology*, 10(2):720, 2014.

# Tests de fonction de répartition cumulative conditionnelle pour l'analyse d'ensembles de gènes de données RNA-seq en cellule unique

Sara Fallet[1,2], Denis Agniel[3], Rodolphe Thiébaut[1,2,4] & Boris Hejblum[1,2]

[1] *Univ. Bordeaux, INSERM, INRIA, Bordeaux Population Health, SISTM team, U1219, F-33000 Bordeaux, France*
[2] *Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France*
[3] *Rand Corporation, Santa Monica, CA 90401, USA*
[4] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*

*sara.fallet@u-bordeaux.fr, dagniel@rand.org*
*rodolphe.thiebaut@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr*

**Résumé.** La technologie de séquençage d'ARN en cellule unique (scRNA-seq) mesure l'expression génique dans des centaines, voire des milliers de cellules à partir d'un seul échantillon biologique, permettant d'étudier les mécanismes moléculaires à l'échelle de la cellule. En immunologie, cette technologie est de plus en plus utilisée pour étudier la réponse immunitaire lors d'une infection (ou vaccination) tout en tenant compte de l'hétérogénéité cellulaire dans le sang. L'analyse de l'expression différentielle identifie les gènes dont l'expression change selon les différentes conditions d'étude. Cependant, les méthodes d'analyse différentielle manquent de puissance statistique et de stabilité, notamment en raison de la nature très dynamique de l'expression génique, de l'hétérogénéité de l'état cellulaire et des limitations technologiques telles que la profondeur de séquençage. En s'intéressant plutôt à des ensembles de gènes associés à des fonctions immunitaires spécifiques, définis à partir de connaissances biologiques *a priori*, on améliore la puissance statistique et la stabilité de l'analyse tout en facilitant l'interprétation biologique des résultats. Nous présentons ici une nouvelle méthode d'analyse différentielle par groupes de gènes adaptée aux données scRNA-seq. Cette méthode repose sur une estimation suivie d'un test de la fonction de répartition conditionnelle de l'expression des gènes au sein d'un groupe. Cette nouvelle méthode s'affranchit ainsi du besoin de faire une hypothèse distributionnelle (délicate pour les données scRNA-seq). Elle est également capable d'analyser des plans expérimentaux complexes, testant l'association de chaque ensemble de gènes avec une ou plusieurs variables d'intérêt (qu'elles soient continues ou discrètes), tout en ajustant éventuellement sur d'autres covariables, dépassant ainsi le cadre usuel de la comparaison simple entre deux groupes. Nous appliquons cette nouvelle méthodologie à deux jeux de données réelles de scRNA-seq étudiant la réponse immunitaire à l'infection par le SARS-CoV-2 chez l'homme, avec respectivement 84 140 cellules T CD8+ provenant de 38 patients et 1 191 463 cellules mononucléaires du sang périphérique provenant de 222 donneurs.

**Mots-clés.** séquençage d'ARN en cellule unique, analyse par groupes de gènes, analyse d'expression différentielle, test d'indépendance conditionnel, fonction de répartition cumulative conditionnelle

**Abstract.** Single-cell RNA-seq (scRNA-seq) technology measures gene expression in hundreds or even thousands of cells from a single biological sample, allowing to study molecular mechanisms at the single-cell resolution. In immunology, this technology is increasingly used to disentangle the complex immune response to infection (or vaccination) while accounting for cellular heterogeneity in the blood. Differential Expression Analysis (DEA) allows to identify which genes are differentially expressed across different conditions, cell types, timings or exposures. However, DEAs often encounter challenges related to statistical power and stability, notably due to the dynamic nature of gene expression and cellular state heterogeneity. Investigating instead gene sets associated with specific immune functions, derived from prior biological knowledge, can enhance the statistical power and stability of the analysis while facilitating the biological interpretation of results.
We introduce a novel gene set analysis method tailored for scRNA-seq data. This method relies on the estimation and testing of conditional distribution functions, eliminating the need for distributional assumptions. This new method is suitable for complex experimental designs, testing the association of each gene set with one or multiple variables of interest (whether continuous or discrete), while potentially adjusting for additional covariates. We apply this new methodology to two single-cell RNA-seq real dataset investigating the immune response to SARS-CoV-2 infection in humans, with respectively 84,140 virus-reactive CD8+ T cells from 38 patients; and 1,191,463 peripheral blood mononuclear cells from 222 donors.

**Keywords.** single cell RNA-seq, gene set analysis, differential expression analysis, conditional independence test, conditional cumulative distribution function

# 1 Introduction

Le séquençage d'ARN en cellule unique permet une analyse plus fine de l'expression génique, permettant la mesure de l'expression génique de chaque cellule individuellement. Comparé au séquençage de l'ARN en masse (*bulk*), qui traite les cellules comme un mélange homogène sans les distinguer, le scRNA-seq offre une résolution plus précise en tenant compte des différences entre les types cellulaires et leurs états. Les données à cellule unique présentent des caractéristiques différentes des données de séquençage d'ARN en masse qui nécessitent une considération spéciale pour le développement d'outils DEA. Notamment, les données scRNA-seq affichent de grandes proportions de zéros, en raison de processus biologiques ou de limitations techniques. Ces données sont également multi-échelles puisqu'elles ont une résolution au niveau des cellules, des échantillons et des conditions. Comme l'analyse des données scRNA-seq présente des particularités uniques, de nouvelles méthodes statistiques sont nécessaires. Les méthodes existantes font des hypothèses fortes sur la distribution des données qui sont très difficiles à vérifier en pratique. Il est donc important de développer des méthodes générales et flexibles pour analyser les données scRNA-seq qui ne nécessitent pas de fortes hypothèses paramétriques.

L'analyse différentielle de l'expression génique permet d'identifier les gènes dont l'expression change entre différentes conditions d'étude (par exemple types cellulaires, temps ou expositions). Un gène est ainsi appelé différentiellement exprimé si son expression est significativement associée aux variations d'un facteur d'intérêt. La plupart des méthodes

existantes pour l'analyse différentielle de données scRNA-seq permettent seulement la comparaison entre deux groupes. En pratique, on peut s'intéresser à des schémas expérimentaux plus complexes, qui nécessitent la comparaison de plus de deux groupes – comme par exemple plusieurs niveau de sévérités d'une maladie.

Nous proposons une nouvelle méthode basée sur les travaux de Gauthier et al. (2021), `citcdf gsa`, qui est non paramétrique et permet de comparer plus de 2 conditions. Elle s'appuie sur un test d'indépendance conditionnel entre l'expression d'un gène $Y$ et une variable d'intérêt $X$, ajusté sur des covariables $Z$. Ce test estime la fonction de répartition conditionnelle (à l'aide de multiples régressions logistiques) afin d'évaluer la potentiel différence selon le conditionnement sur $X$. Les performances de notre méthode ont été étudiées par des simulations numérique ainsi que pour l'analyse de deux jeux de données réelles scRNA-seq étudiant chacun la réponse immunitaire à l'infection par le SARS-CoV-2.

# 2 Méthode

## 2.1 `citcdf` : analyse gène par gène (Gauthier et al., 2021)

On effectue un test d'indépendance conditionnel entre l'expression d'un gene $Y$ et un facteur d'intérêt $X$, sachant des covariables $Z$ :

$$H_0 : Y \perp X \mid Z,$$

où $X$ et $Z$ peuvent être respectivement le statut d'une maladie ou le bras d'un vaccin, et le type d'une cellule ou l'âge du patient par exemple.

Notre méthode d'analyse d'expression différentielle se base sur l'estimation de fonction de répartition conditionnelle. Ainsi l'hypothèse nulle de notre test peut-être reformulée comme :

$$H_0 : F_{Y|X,Z}(y, x, z) = F_{Y|Z}(y, z),$$

où $F_{Y|X,Z}(y, x, z)$ est la fonction de répartition conditionnelle de $Y$ sachant $X$ et $Z$, et $F_{Y|Z}(y, z)$ est la fonction de répartition « marginale » (au sens indépendemment de $X$) de $Y$ (mais toujours conditionnellement à $Z$). On teste l'égalité entre ces deux fonctions car, si un groupe de facteur est associé à l'expression d'un gène, alors la $F_{Y|X,Z}$ sera significativement différente de $F_{Y|Z}$.

La fonction de répartition conditionnelle peut s'exprimer sous la forme d'une probabilité :

$$F_{Y|X,Z}(y, x, z) = \mathbb{P}(Y \leq y | X = x, Z = z).$$

Pour chaque gène, on peut ainsi choisir une séquence de $p$ seuils réguliers et ordonnés $\omega_j$ où $j = 1, \ldots, p$. Ainsi $F_{Y|X,Z}$ peut s'écrire comme l'espérance conditionnelle d'un variable binaire :

$$F_{Y|X,Z}(y, x, z) = \mathbb{E}(\mathbb{1}_{\{Y_i \leq \omega_j\}} | X_i = x, Z_i = z) \tag{1}$$

On peut alors estimer (1) par un séquence de $p$ régressions logistiques :

$$g(\mathbb{E}[\mathbb{1}_{\{Y_i \leq \omega_j\}} | X_i, Z_i]) = \beta_{0j} + \boldsymbol{\beta}_{1j}\mathbf{X_i} + \boldsymbol{\beta}_{2j}\mathbf{Z_i},$$

où $i = 1, \ldots, n$ indexe les observations de chaque cellule, $\boldsymbol{\beta}_{1j} = (\beta_{1j1}, \ldots, \beta_{1js_1})$ quantifie l'influence de $X_i$ sur $\mathbb{P}(Y_i \leq \omega_j)$ et $\boldsymbol{\beta}_{2j} = (\beta_{2j1}, \ldots, \beta_{2js_2})$ quantifie l'influence de $Z_i$ sur $\mathbb{P}(Y_i \leq \omega_j)$. $s_1$ et $s_2$ désignent respectivement le nombre total de conditions et de variables considéré.

Si le facteur d'intérêt $X$ n'a aucun lien avec l'expression du gène $Y$, $\boldsymbol{\beta}_{1j}$ sera égal à 0. De ce fait l'hypothèse nulle de notre test devient :

$$H_0 : \boldsymbol{\beta}_1 = 0,$$

où $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{1j}, \ldots, \boldsymbol{\beta}_{1p})$ est la matrice des $\beta_1$ pour chaque seuil d'expression choisi et pour chaque condition testée.

La statistique de test du score associée peut s'estimer comme $\widehat{D}_n = n \sum_{j=1}^{p} \sum_{k=1}^{s_1} \widehat{\beta}_{1jk}^2$, ainsiq que sa distribution asymptotique $\widehat{D}_n \underset{n \to +\infty}{\longrightarrow} \sum_{q=1}^{ps_1} \widehat{a}_q \chi_1^2$ avec les $\widehat{\beta}_{1jk}$ qui sont approchés grâce à la méthode des moindres carrés ordinaires (afin d'accélérer les calculs), tandis que les $\widehat{a}_q$ sont les valeurs propres de la matrice de variance covariance des $\beta_{1jk}$ notée $\Sigma$. Enfin, on calcule les p-valeurs en comparant la statistique de test observée $\widehat{D}_n$ avec sa distribution asymptotique.

## 2.2 `citcdf gsa` : analyse par groupe de gènes

Dans le cadre de l'analyse par groupe de gènes, on estime cette fois (1) avec le modèle :

$$g(\mathbb{E}[\mathbb{1}_{\{Y_i \leq \omega_j\}} | X_i, Z_i]) = \beta_{0jl} + \boldsymbol{\beta}_{1jl} \mathbf{X_i} + \boldsymbol{\beta}_{2jl} \mathbf{Z_i},$$

où $l = 1, \ldots, m$ désigne les différents gènes du groupe de gènes.

Notre hypothèse nulle devient naturellement $H_0 : \boldsymbol{\beta}_1 = 0$, avec $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{111}, \ldots, \boldsymbol{\beta}_{1p1}, \ldots, \boldsymbol{\beta}_{11m}, \ldots, \boldsymbol{\beta}_{1pm})$ la matrice des $\beta_{1jlk}$ pour chaque seuils, de chaque gène du groupe de gène. Ainsi la statistique de test et sa distribution asymptotique deviennent :

$$\widehat{S}_n = n \sum_{j=1}^{p} \sum_{k=1}^{s_1} \sum_{l=1}^{m} \widehat{\beta}_{1jkl}^2 \qquad \widehat{S}_n \underset{n \to +\infty}{\longrightarrow} \sum_{j=1}^{ps_1 m} \widehat{a}_j \chi_1^2$$

Lors du calcul de la matrice de variance covariance $\widehat{\Sigma}$, qui permet d'estimer les valeurs propres $\widehat{a}_j$, on a des termes croisés supplémentaires entre les $\widehat{\boldsymbol{\beta}}_{1jkl}$ de chaque gène du groupe set.

# 3 Résultats

## 3.1 Simulations numériques

Afin d'évaluer les performances du test proposé dans `citcdf gsa`, nous avons généré des simulations numériques gaussiennes, paramétrées avec $n = 100$ cellules et $r = 200$ gènes. Des facteurs d'intérêt $X_1$ et $X_2$, et des covariables $Z_1$ et $Z_2$ ont été générés à partir de distributions normales, et finalement la matrice $Y$ d'expression des gènes a également été générées grâce à l'addition d'un bruit gaussien. Les résultats testant des groupes de deux gènes indépendant ont été agrégés sur 1500 simulations, et sont représentés en Figure 1. On constate un controle adéquate de l'erreur de Type-I sous $H_0$, accompagné d'une puissance adéquate sous $H_1$.

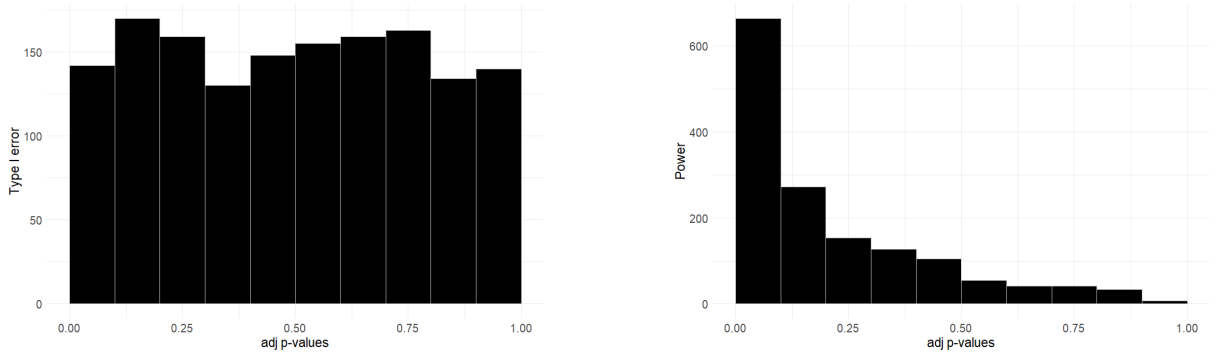



FIGURE 1 : Performance du test en simulation. À gauche : test sous l'hypothèse nulle $H_0$. À droite : test sous l'hypothèse alternative $H_1$

## 3.2 Applications à des données scRNA-seq réelles

Kusnadi et al. (2021) ont mesuré l'expression de 13 816 gènes au sein de 84 140 cellules T CD8 positives, réactives au virus SARS-CoV-2, chez 39 patients atteints de la COVID-19 répartis en trois groupes : 17 patients non hospitalisés, 13 hospitalisés et 9 hospitalisés en unité de soins intensifs. Kusnadi et al. (2021) ont mis en évidence l'importance de 6 groupes de gènes *Exhaustion Consensus*, *Genes upregulated in cytotoxicity*, *Type I and II Inferon signaling genes*, *Viral response*, *Hallmark Glycolysis* et *Genes upregulated in cell cycle* contenant entre 11 et 226 gènes. `citcdf gsa` identifie l'ensemble de ces 6 groupes de gènes comme significatif avec un taux de fausse découverte inférieur à 5%. La Figure 2 représente l'expression des gènes du groupe de gènes annotés *Genes upregulated in cytotoxicity* regroupé selon les 3 sévérités cliniques.

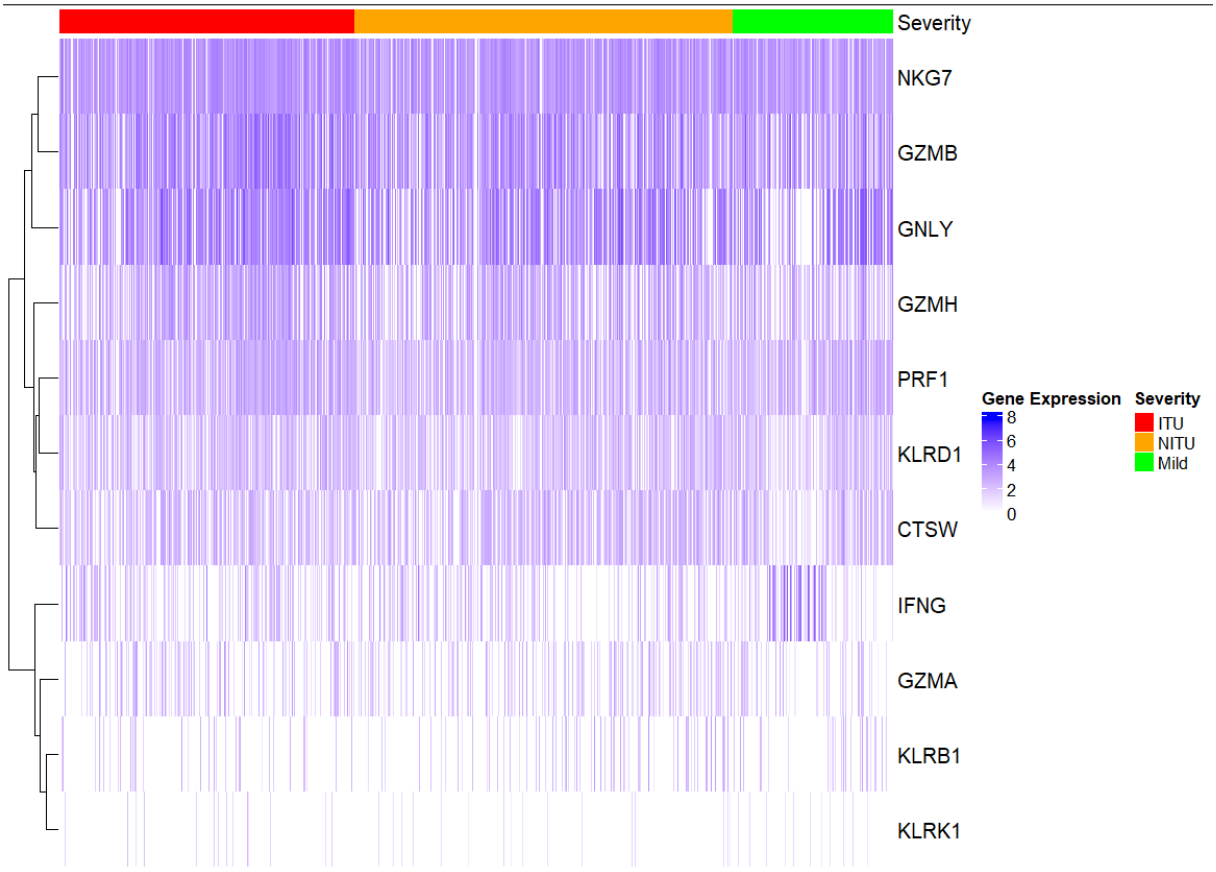



FIGURE 2 : Heatmap de l'expression du groupe de gènes *Genes upregulated in cytotoxicity*

# 4 Discussion

`citcdf gsa` est une méthode d'analyse d'expression différentielle par groupe de gène applicable aux données scRNA-seq. Elle réalise un test d'indépendance conditionnel qui se base sur l'estimation de la fonction de répartition conditionnelle grâce à de multiple régressions logistiques. Elle permet de prendre en compte des schémas expérimentaux complexes, notamment en permettant d'étudier plus de 2 conditions expérimentales ou en permettant d'ajuster sur des covariables. Étant une méthode non-paramétrique, `citcdf gsa` peut suivre n'importe quelle méthode de normalisation des données de séquençage. `citcdf gsa` présente de bonnes performances en simulation avec un contrôle efficace de l'erreur de type I et une puissance statistique raisonnable.

Nous avons implémenté cette méthode dans le package R `citcdf`, branche `gsa` pour l'analyse par groupe de gènes, disponible sur *GitHub* : `https://github.com/sistm/citcdf/tree/gsa`.

# Références

Gauthier, M., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2021). Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis. Preprint, Bioinformatics.

Kusnadi, A., Ramírez-Suástegui, C., Fajardo, V., Chee, S. J., Meckiff, B. J., Simon, H., Pelosi, E., Seumois, G., Ay, F., Vijayanand, P., and Ottensmeier, C. H. (2021). Severely ill patients with COVID-19 display impaired exhaustion features in SARS-CoV-2–reactive CD8 $^{+}$ T cells. *Science Immunology*, 6(55):eabe4782.

# Procédure de test hiérarchique pour l'analyse différentielle de données Hi-C

Élise Jorge[1,2], Pierre Neuvial[3], Nathalie Vialaneix[2] & Sylvain Foissac[1]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France, {elise.jorge, sylvain.foissac}@inrae.fr*

[2] *Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet-Tolosan, France, nathalie.vialaneix@inrae.fr*

[3] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France, pierre.neuvial@math.univ-toulouse.fr*

**Résumé.** Les données Hi-C fournissent une information sur l'organisation tridimensionnelle du génome à partir de mesures d'interactions entre positions génomiques le long de la chromatine. Cette structure en trois dimensions a un rôle important dans la régulation de l'expression des gènes. L'objectif de l'analyse différentielle est d'identifier, à partir de réplicats obtenus dans deux conditions biologiques différentes, des régions génomiques qui présentent des différences significatives de structure entre les deux conditions. Ici, nous proposons de nous appuyer sur une modélisation hiérarchique des données Hi-C, permettant de tenir compte de la dépendance spatiale présente dans ce type de données. En utilisant un travail précédent permettant de représenter les données par des arbres et de les utiliser pour réaliser des tests, on peut, à partir de régions définies *a priori*, identifier celles qui sont d'intérêt. On s'intéresse ici au développement d'une méthode permettant d'identifier automatiquement de telles régions.

**Mots-clés.** données Hi-C, génomique 3D, arbre binaire, classification hiérarchique, tests multiples

**Abstract.** Hi-C data provide insights into the three-dimensional organization of the genome by measuring interactions between genomic positions along the chromatin. This three-dimensional structure plays a crucial role in regulating gene expression. Differential analysis aims to identify genomic regions that display significant differences in structure between two different biological conditions. Here, we propose a hierarchical modeling approach to analyze Hi-C data, allowing to incorporate the inherent spatial dependence within the data. We build upon a recent method that represents the data as trees and performs targeted tests on predefined regions. The objective is to develop a method that automatically identifies regions of interest, streamlining the differential analysis process.

**Keywords.** Hi-C data, 3D genomics, binary tree, hierarchical classification, multiple testing

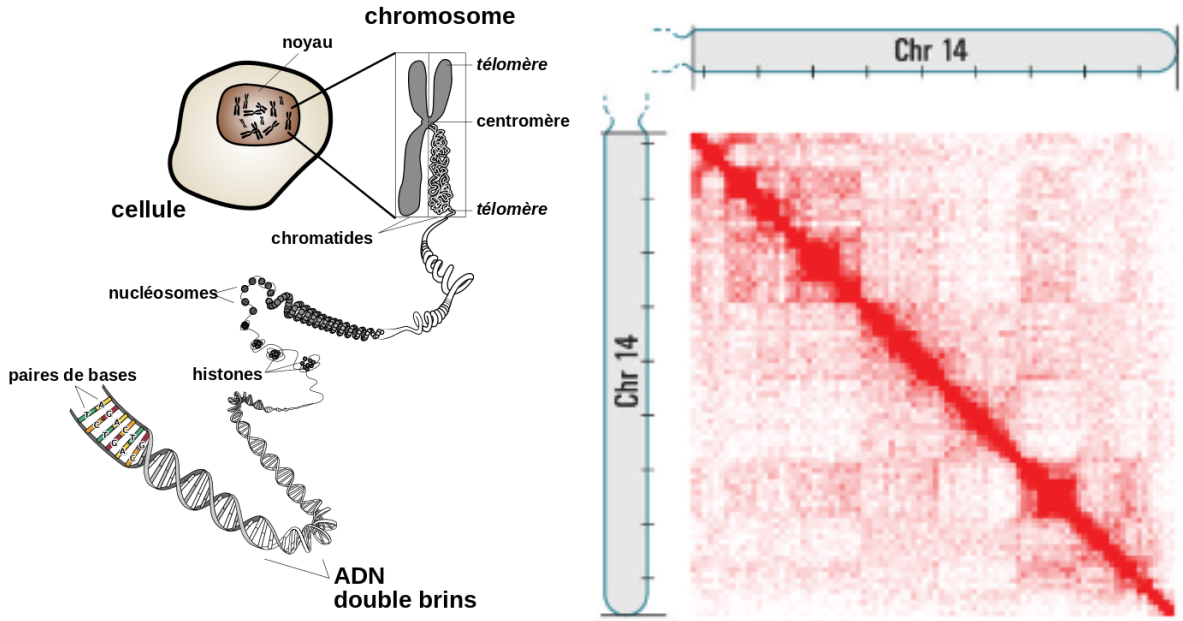



FIGURE 1 – Gauche : Schéma de la compaction de l'ADN en chromosome ("Chromosome fr" par Phrood commonswiki, Wikimedia Commons). Droite : Matrice Hi-C du chromosome 14 de [1].

# 1 Introduction

### Structure de l'ADN et données Hi-C

La chromatine est compactée au sein du chromosome selon une structure hiérarchique, comme illustré sur la figure 1 (gauche). Les données Hi-C sont des données de séquençage haut-débit qui permettent d'obtenir des informations sur l'organisation tridimensionnelle du génome dans la cellule, en mesurant la fréquence d'interactions spatiales entre régions génomiques. L'étude de ces données a permis de montrer qu'il existait, le long de la chromatine, des régions génomiques appelées TADs (*Topologically Associating Domains*) au sein desquels les interactions sont fréquentes.

L'apparition de modifications dans cette structure de compaction, par exemple la disparition d'une frontière entre deux TADs impliquant leur fusion, peut avoir un impact majeur sur l'expression des gènes dans la zone considérée. Ces modifications peuvent provoquer des pathologies neurologiques [2] ou des malformations [3].

D'une manière plus formelle, les données Hi-C se présentent sous la forme d'une matrice carrée symétrique dont l'entrée $(i, j)$ correspond au nombre de contacts observés dans l'expérience Hi-C entre les positions génomiques $i$ et $j$. La figure 1 (droite) représente une telle matrice, dans laquelle l'intensité de couleur est proportionnelle à la valeur du nombre de contacts correspondant.

### Analyse différentielle de données Hi-C

On s'intéresse ici à un problème d'analyse différentielle entre des ensembles de matrices Hi-C obtenues dans deux conditions différentes, $\mathcal{C}_1$ et $\mathcal{C}_2$. L'objectif est d'identifier des régions génomiques qui présentent des différences significatives d'interactions entre ces deux conditions.

Formellement, on possède $r = r_1 + r_2$ matrices de taille $p \times p$ où $M_k^{\mathcal{C}_1}$ $(k = 1, \cdots, r_1)$ (resp. $M_l^{\mathcal{C}_2}$ $(l = 1, \cdots, r_2)$) correspond à la matrice obtenue pour le $k$-ème (resp. $l$-ème) réplicat de la condition $\mathcal{C}_1$ (resp. $\mathcal{C}_2$).

**État de l'art**

Différentes méthodes ont été développées pour répondre à ce problème dont **diffHic** [4], **FIND** [5], **HiCcompare** [6], **multiHiCcompare** [7], **Selfish** [8] et **ACCOST** [9] mais celles-ci ne tiennent pas (ou peu) compte de la structure hiérarchique des données et produisent, de manière indépendante, une $p$-valeur par paire de positions (en particulier, [4, 7, 9] fondent leur approche sur une modélisation des comptages par la loi Binomiale Négative). Les détections positives sont donc fréquemment éparpillées sur l'intégralité de la matrice, sans relation avec une structure fonctionnelle du génome et donc peu interprétables.

La méthode *treediff* [10] permet de prendre en compte les dépendances entre positions génomiques induites par la structure 3D de la chromatine en représentant les matrices par des dendrogrammes qui sont un type particulier d'arbre binaire. À partir de cette représentation hiérarchique des données, des tests individuels position par position sont réalisés. Une méthode d'agrégation des tests individuels permet ensuite, pour une zone de la matrice donnée, d'identifier s'il existe au sein de cette zone au moins une interaction significativement différentielle entre les deux conditions. Comme nous l'expliquons ci-dessous, la limite de cette méthode est qu'elle repose sur une définition préalable des zones à tester par l'utilisateur. C'est cette limite que nous abordons dans ce travail.

## 2 La méthode *treediff*

On présente ici les principales étapes de la méthode *treediff* [10].

1. **Étape 1 : Classification ascendante hiérarchique sous contrainte de contiguïté.**
   Cette première étape a pour but d'obtenir une classification des positions génomiques pour chaque matrice dans chaque condition. Pour cela, la classification hiérarchique ascendante avec contrainte de contiguité (CAHCC) définie dans [11] est utilisée. Cette méthode consiste à appliquer une CAH à noyau, bien adaptée à des données qui sont des similarités. La contrainte de contiguïté dans la classification impose de ne regrouper que des classes et/ou des feuilles adjacentes le long du génome. On obtient, pour chaque matrice, un dendrogramme représentant une classification des données rendant compte de la dépendance spatiale dans les données Hi-C comme illustré sur la figure 2.
2. **Étape 2 : Comparaison de dendrogrammes et tests de Student.**

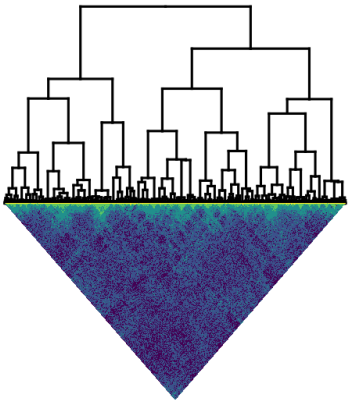


FIGURE 2 – (Demie) matrice Hi-C (en bas) et le dendrogramme associé obtenu par CAHCC (en haut).

Afin de pouvoir comparer les dendrogrammes entre eux et déterminer l'existence de différences d'interactions, des distances cophénétiques sont calculées entre toutes les paires de positions génomiques, $i$ et $j$, pour chaque dendrogramme.
Ainsi, $\forall\, i,j \in \{1,\cdots,p\}$, $k \in \{1,\cdots,r_1\}$ (resp. $l \in \{1,\cdots,r_2\}$), on notera $d_{(i,j)}^{\mathcal{C}_1,k}$ (resp. $d_{(i,j)}^{\mathcal{C}_2,l}$ ) la distance cophénétique entre les positions $i$ et $j$ pour le $k$-ème (resp. $l$-ème) réplicat de la condition $\mathcal{C}_1$ (resp. $\mathcal{C}_2$). Ensuite, pour deux positions génomiques $i$ et $j$, l'hypothèse « *Il n'existe pas de différence d'interaction entre les conditions* $\mathcal{C}_1$ *et* $\mathcal{C}_2$ *pour les positions* $i$ *et* $j$ » est formulée sous la forme d'une hypothèse nulle sur l'égalité des moyennes de la distance cophénétique entre $i$ et $j$ pour les conditions $\mathcal{C}_1$ et $\mathcal{C}_2$, $H_0^{(i,j)}$ : « $\mu_{ij}^{\mathcal{C}_1} = \mu_{ij}^{\mathcal{C}_2}$ ». Cette partie conduit à l'obtention d'une $p$-valeur $\pi_{i,j}$ pour chaque paire de positions $(i,j)$.

3. **Étape 3 : Agrégation de Simes.**
   La dernière partie consiste à utiliser les $p$-valeurs individuelles pour réaliser un test de l'hypothèse $H_0^C = \cap_{i,\,j\in C,\ i<j} H_0^{(i,j)}$, où $C$ est un intervalle de positions génomiques contigües. Pour ce faire, les $\pi_{i,j}$ sont agrégées par la méthode de Simes [12]. De manière plus précise, si $n = \frac{|C|(|C|-1)}{2}$ est le nombre de paires de positions génomiques dans $C$, l'agrégation de Simes consiste à calculer une $p$-valeur corrigée :

$$\pi_{\text{Simes}}^C = \min\left\{ n \times \frac{\pi_{(k)}}{k};\ k = 1,\cdots,n \right\}, \tag{1}$$

   où $\pi_{(k)}$ est la $k$-ème plus grande $p$-valeur parmi les $(\pi_{i,j})_{i,j\in C,\ i<j}$.
   Sous des hypothèses de dépendance positive de type PRDS (Positive Regression Dependency on a Subset) [13] sur les $p$-valeurs individuelles, l'erreur de type I est contrôlée au niveau $\alpha$, càd : $\mathbb{P}_{H_0^C}(\pi_{Simes}^C \leq \alpha) \leq \alpha$. L'hypothèse PRDS est classique en tests multiples et considérée comme réaliste en génomique [14].

En résumé, la méthode *treediff* permet d'inférer l'existence d'au moins une interaction différentielle au sein d'une zone prédéfinie de la matrice. Cette méthode requiert donc une connaissance préalable des données afin de choisir des zones à tester. De plus, lorsque plusieurs zones sont testées, une correction de tests multiples est nécessaire.

# 3 Une méthode « *data-driven* » d'analyse différentielle de données Hi-C

L'objectif de cette section est de présenter comment, à partir de la méthode *treediff*, nous proposons une méthode qui utilise les données pour déterminer quelles zones de la matrice tester tout en conservant le contrôle de l'erreur. Pour cela, nous avons considéré plusieurs méthodes de tests multiples sur des données structurées que nous présentons dans la section 3.1 avant de décrire notre proposition dans la section 3.2.

## 3.1 Tests multiples dans un cadre hiérarchique

La plupart des méthodes de contrôle de tests multiples exploitant une structure hiérarchique ont été développées dans un cadre de sélection de variables ou d'un autre type de test portant sur des variables $(X^j)_{j=1,\dots,p}$ dans lequel on connaît une structure hiérarchique entre les $X^j$ (généralement un dendrogramme obtenu à partir d'une classification ascendante hiérarchique des $X^j$). Dans ce cadre, on connaît la $p$-valeur $\phi_j$ associé à chaque $X^j$ indépendamment et contrôlant l'hypothèse $\mathrm{H}_0^j$ portant sur $X^j$. On s'intéresse alors à trouver les groupes (classes) de variables, $C$, rejetant l'hypothèse nulle $H_0^C = \cap_{j\in C} H_0^j$. Ce cadre diffère donc un peu de celui que nous avons décrit précédemment puisqu'il ne considère que les hypothèses nulles liées directement aux variables, et pas l'ensemble des hypothèses nulles liées aux interactions entre ces variables.

Nous définissons deux types d'erreurs classiques pour le cadre des tests multiples : le FWER (*Family Wise Error Rate*) [15] et le FDR (*False Discovery Rate*) [16]. Soit $\mathcal{H}$ un ensemble d'hypothèses et $\mathcal{H}_0 \subseteq \mathcal{H}$ le sous-ensemble des hypothèses vraies. Soit $R$ l'ensemble des hypothèses rejetées par la procédure de test et $|R \cap \mathcal{H}_0|$ le nombre de faux positifs. Le FDR est défini par $\mathbb{E}(\frac{|R\cap\mathcal{H}_0|}{|R|\vee 1})$ qui est l'espérance de la proportion de faux positifs parmi les hypothèses rejetées. Le FWER correspond à $\mathbb{P}(|R\cap\mathcal{H}_0| \geq 1)$, la probabilité de faire au moins une erreur de type I.

Une des premières méthodes permettant d'aborder la question du contrôle de l'erreur dans un contexte de données structurées hiérarchiquement est celle décrite dans [17] qui propose une procédure de test d'hypothèses structurées en un unique arbre et assurant un contrôle du FDR. Toutefois, comme discuté dans [14], dans le cadre de tests portant sur un cluster d'hypothèses, le contrôle du FDR ne donne pas une garantie suffisante sur les résultats.

Une approche permettant de contrôler le FWER dans un cadre similaire a été proposée dans [18]. Cette méthode s'applique à des données organisées hiérarchiquement en arbre et consiste à parcourir l'arbre de haut en bas afin d'identifier les plus petits clusters rejetant l'hypothèse nulle tout en assurant un contrôle global du FWER sur tout le parcours de l'arbre. Cette méthode a été proposée dans le cadre de la sélection de variables, en particulier dans le cas où celles-ci sont fortement corrélées.

Depuis, d'autres approches similaires ont été proposées, comme celle de [19], qui est une méthode de test pour des données organisées en DAG (*Directed Acyclic Graph*). Celle-ci est

appliquée à un graphe d'ontologie de gènes. Un dendrogramme pouvant être vu comme un type de DAG, la structure de données utilisée dans cette méthode est compatible avec le cas d'application de *treediff*. Cependant, une des hypothèses utilisées par les auteurs est que, pour trois clusters $A$, $B$ et $C$ tels que $A = B \cup C$, alors si $H_0^A$ et $H_0^B$ sont vraies, $H_0^C$ est vraie aussi. Ceci n'est pas vérifié dans notre cadre car si aucune interaction différentielle n'est trouvée dans le cluster A et le cluster B, il pourrait quand même y avoir une interaction différentielle entre une position $i$ de A et une position $j$ de B donc l'hypothèse $H_0^C$ ne serait pas vérifiée. Cette méthode n'est donc pas adaptée à notre cadre d'étude.

Enfin, une procédure de test pour des données ordonnées en temps ou en espace est proposée dans [20]. La méthode consiste à représenter les données sur un graphe puis à utiliser une procédure de test le long de ce graphe qui permet un contrôle du FWER. Bien que les hypothèses que l'on souhaite tester soient organisées spatialement, le fait que cette méthode ne propose pas d'utiliser de structure hiérarchique *a priori* sur les données rend difficile son adaptation au contexte des données Hi-C.

Nous avons donc choisi de nous appuyer sur le cadre formel de [18], bien adapté au contexte des données Hi-C, et de l'étendre au cas particulier où l'hypothèse nulle correspond à une intersection d'hypothèses nulles sur des interactions.

## 3.2 Contrôle de tests multiples pour l'analyse de données Hi-C

Pour adapter l'approche de [18] au cadre décrit dans la section 2, nous avons besoin d'une structure hiérarchique unique pour appuyer le parcours des tests à effectuer. Aussi, en complément des CAHCC réalisées sur chacune des matrices Hi-C individuelles, nous proposons de construire une CAHCC « consensuelle » basée sur la matrice $M = \sum_{k=1}^{r_1} M_k^{\mathcal{C}_1} + \sum_{k=1}^{r_2} M_k^{\mathcal{C}_2}$.

Le parcours du dendrogramme se fait dans le sens inverse de l'agrégation des classes (c'est-à-dire, de la racine du dendrogramme, jusqu'aux feuilles, en parcourant chaque nœud du dendrogramme, correspondant à une classe, avant de parcourir ses nœuds enfants). Pour chaque classe $C$ rencontrée dans le parcours du dendrogramme, on dispose de la $p$-valeur obtenue par la méthode *treediff*, $\pi^C$ : cette $p$-valeur contrôle la probabilité de rejet à tort de $C$ sous l'hypothèse $H_0^C$. Afin d'assurer un contrôle du FWER sur l'ensemble de l'arbre, on réalise alors, pour chaque classe, deux ajustements de cette $p$-valeur, un premier relatif à la taille de la classe considérée et un second relatif à la hiérarchie :

1. on appelle *p-valeur ajustée de* $C$ la quantité :

$$\pi_{adj}^C = \frac{p}{|C|}\pi^C \tag{2}$$

   Ainsi, plus une classe est petite (et donc plus elle est « lointaine » dans le parcours de la hiérarchie), plus elle est pénalisée ;

2. on appelle *p-valeur hiérarchiquement ajustée* la quantité :

$$\pi_{adj,h}^C = \max_{D, D \supset C} \pi_{adj}^D \tag{3}$$

Ainsi, les classes rejetées au seuil $\alpha$ par cette procédure sont des classes dont tous les parents ont été rejetés. Réciproquement, cet ajustement hiérarchique permet de s'assurer que si une classe n'est pas rejetée, ses descendantes ne le sont pas non plus. La procédure est illustrée dans la figure 3. Enfin, nous avons prouvé que cette approche permet bien un

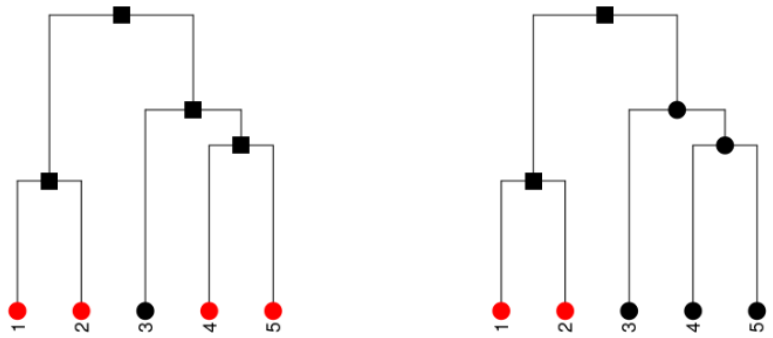



FIGURE 3 – Illustration de la procédure globale de parcours du dendrogramme consensus. Les classes rejetées sont représentées par un carré et les classes non-rejetées sont représentées par un cercle. Les feuilles (qui, dans le cas de l'approche *treediff*, ne peuvent être testées) sont représentées par un cercle. Les feuilles qui correspondent aux positions génomiques identifiées par la procédure sont représentées en rouge.

contrôle du FWER global, comme dans le cadre décrit par [18] (résultat non montré ici).

# 4 Application, résultats et perspectives

Nous avons implémenté la méthode décrite en section 3.2, et l'avons testée sur des données Hi-C issues de lignées cellulaires murines [21] pour deux conditions biologiques qui correspondent à des stades de différenciation cellulaire différents de cellules neuronales.

L'analyse des premiers résultats nous permet de faire deux remarques principales : en premier lieu, on observe ce qui est appelé le *spatial specificity paradox* et qui n'est pas spécifique à notre contexte applicatif. Ce paradoxe postule que les plus grosses classes rejetées de la hiérarchie ne sont pas les plus informatives. En second lieu, on observe que le fait que les tests sont, dans le cas des données Hi-C, des tests d'interactions modifie l'interprétation des résultats.

Les dendrogrammes de la figure 4 nous permettent d'illustrer ces deux remarques.

Dans le dendrogramme de gauche, la classe qui contient toutes les positions est rejetée et ses deux classes descendantes ne le sont pas. C'est donc la région génomique correspondant à l'ensemble des positions qui est identifiée. Or, l'information que nous donne la procédure est qu'au sein des classes $\{1, 2, 3, 4, 5, 6, 7\}$ et $\{8, 9, 10\}$, il n'existe pas d'interaction différentielle. Le rejet de la classe contenant toutes les positions signifie donc qu'il existe probablement au moins une interaction différentielle entre une position de la classe $\{1, 2, 3, 4, 5, 6, 7\}$ et une

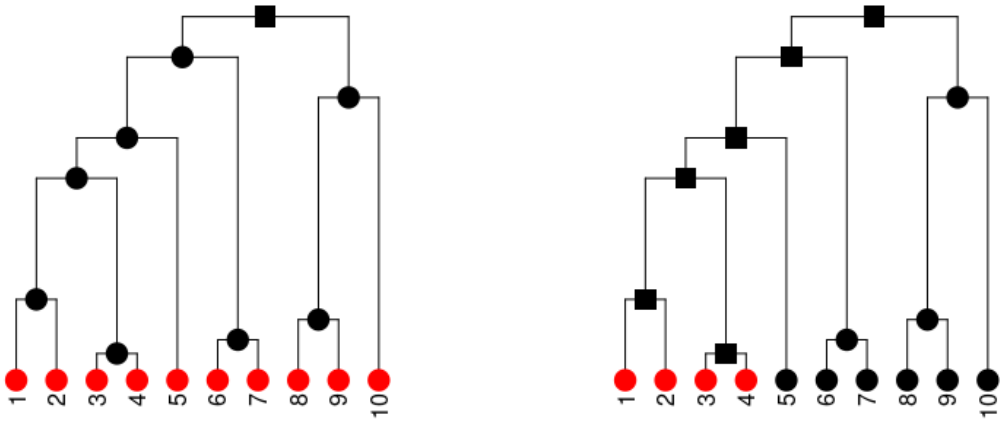



FIGURE 4 – Deux dendrogrammes illustrant des spécificités observées dans les résultats. Les classes rejetées sont représentées par un carré et les classes non-rejetées sont représentées par un cercle. Les feuilles ne pouvant être testées, elles sont représentées par un cercle. Les feuilles qui correspondent aux positions génomiques identifiées par la procédure sont représentées en rouge.

position de la classe $\{8, 9, 10\}$. Cet exemple illustre que, pour une région identifiée de taille significative, on ne possède pas d'information sur le nombre d'interactions différentielles ni sur leur localisation précise.

Sur le dendrogramme de droite, la classe contenant toutes les positions est également rejetée. Néanmoins, comme la classe $\{1, 2, 3, 4, 5\}$ est, cette fois-ci, aussi rejetée, ce n'est pas la classe globale qui est identifiée par la méthode. Ainsi, bien que des interactions différentielles puissent exister entre les classes $\{1, 2, 3, 4, 5, 6, 7\}$ et $\{8, 9, 10\}$, cette information est ignorée dans la suite de l'exploration de la hiérarchie. Sur ce même dendrogramme, on peut également pointer une limite de la méthode liée au fait de descendre au maximum le long de la hiérarchie : les classes $\{1, 2\}$ et $\{3, 4\}$ sont identifiées par la méthode mais de manière disjointe. En effet, on déduit de la procédure de test qu'il existe au moins une interaction différentielle dans $\{1, 2\}$ et de même dans $\{3, 4\}$. Cependant, bien que la classe $\{1, 2, 3, 4\}$ ait été rejetée par la méthode, on ne sait pas s'il existe des interactions différentielles entre les classes $\{1, 2\}$ et $\{3, 4\}$.

Ces exemples illustrent certaines limites de la méthode et démontrent la nécessité d'un travail supplémentaire d'analyse et de représentation des résultats afin de tirer un maximum d'information de la procédure de test et d'obtenir des régions génomiques d'intérêt d'un point de vue biologique. Les perspectives envisagées en ce sens ainsi qu'une analyse plus complète des résultats seront discutées lors de la présentation.

**Remerciements**

Ce travail est soutenu par le groupe de travail ChrocoNET financé par le métaprogramme INRAE DIGIT-BIO. La thèse d'Élise Jorge est financée par INRAE.

# Bibliographie

[1] Erez Lieberman-Aiden, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M.A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[2] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3d genome. *Nature Reviews Genetics*, 19(7):453–467, 2018.

[3] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[4] Aaron T.L. Lun and Gordon K. Smyth. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16:258, 2015.

[5] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q. Zhang. FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research*, 28:412–422, 2018.

[6] John C. Stansfield, Kellen G. Cresswell, Vladimir I. Vladimirov, and Mikhail G. Dozmorov. HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics*, 19:279, 2018.

[7] John C. Stansfield, Kellen G. Cresswell, and Mikhail G. Dozmorov. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics*, 2019. Forthcoming.

[8] Abbas Roayaei Ardakany, Ferhat Ay, and Stefano Lonardi. Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics*, 35(14):i145–i153, 2019.

[9] Kate B. Cook, Borislav H. Hristov, Karine G. Le Roch, Jean-Philippe Vert, and William Stafford Noble. Measuring significant changes in chromatin conformation with ACCOST. *Nucleic Acids Research*, 48(5):2303–2311, 2020.

[10] Pierre Neuvial, Nathanaël Randriamihamison, Marie Chavent, Sylvain Foissac, and Nathalie Vialaneix. A two-sample tree-based test for hierarchically organized genomic signals. *Journal of the Royal Statistical Society, Series C*, 2024. Forthcoming.

[11] Nathanaël Randriamihamison, Nathalie Vialaneix, and Pierre Neuvial. Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. *Journal of Classification*, 38:363–389, 2021.

[12] R. John Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[13] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[14] Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, November 2011. Publisher: Institute of Mathematical Statistics.

[15] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[17] Daniel Yekutieli. Hierarchical false discovery rate – Controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008.

[18] Nicolai Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008.

[19] Rosa J. Meijer and Jelle J. Goeman. A multiple testing method for hypotheses structured in a directed acyclic graph. *Biometrical Journal*, 57(1):123–143, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201300253.

[20] Rosa J. Meijer, Thijmen J. P. Krebs, and Jelle J. Goeman. A region-based multiple testing method for hypotheses ordered in space or time. *Statistical Applications in Genetics and Molecular Biology*, 14(1):1–19, 2015.

[21] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572.e24, 2017.

# Prediction of gene expression using whole-genome epigenomic signals

Mathilde Bruguet[1] & Romane Leroy[2] & Santa Kirezi[3] & Romain Tavenard[4] & Magalie Houée-Bigot[5] & Gaël Le Trionnaire[6] & Nadia Ponts[7] & David Causeur[8]

[1] *Institut Agro, Rennes, France, mathilde.bruguet@agrocampus-ouest.fr*
[2] *Institut Agro, Rennes, France, romane.leroy@agrocampus-ouest.fr*
[3] *Institut Agro, Rennes, France, santa.kirezi@agrocampus-ouest.fr*
[4] *LETG - UMR CNRS 6554, France, romain.tavenard@univ-rennes2.fr*
[5] *Institut Agro, Rennes, France, magalie.houee@agrocampus-ouest.fr*
[6] *IGEPP - UMR INRAE 1349, France, gael.le-trionnaire@inrae.fr*
[7] *MycSA - UR INRAE 1264, France, nadia.ponts@inrae.fr*
[8] *IRMAR - UMR CNRS 6625, France, david.causeur@institut-agro.fr*

**Résumé.** Pour survivre et se développer, les agents phytopathogènes doivent s'adapter à une variété de stress environnementaux. A des horizons de temps courts, les variations épigénétiques peuvent permettre des réponses phénotypiques adaptatives en modifiant les réseaux d'expression génique, sans changement de séquence génomique. Ainsi, les génomes et les épigénomes interagissent avec l'environnement et contribuent à l'apparition de nouveaux phénotypes, dont l'adaptabilité est une caractéristique clé de la résilience. Si l'adaptation des espèces par variation génétique fait l'objet d'intenses recherches, les composantes épigénétiques de l'innovation phénotypique restent peu étudiées notamment du fait de la difficulté de s'affranchir de variations du fond génétique entre les générations. Dans ce contexte, les organismes se reproduisant de manière clonale sont d'excellents modèles pour étudier la contribution de l'épigénétique dans les processus adaptatifs. Parmi ceux-ci, le champignon filamenteux phytopathogène producteur de mycotoxines *Fusarium graminearum* est un bioagresseur de grandes cultures aux capacités de résilience remarquables.

Les variations d'accessibilité de la chromatine par repositionnement des nucléosomes sont des mécanismes épigénétiques clés qui modulent l'expression des gènes. Leur étude est possible grâce à des méthodes de séquençage à haut débit (MAINE-seq) générant des signaux complexes et hétérogènes dont la valorisation par les méthodes actuelles de la statistique génomique est limitée. L'apprentissage statistique des motifs d'association entre un signal épigénomique observé pour chaque gène sur une région large du génome incluant sa partie codante et sa région promotrice et l'expression de ce gène mesurée par séquençage de l'ARN (RNA-seq) offre des perspectives pour comprendre les capacités d'un organisme à s'adapter à un stress par des mécanismes épigénétiques. Or, les études comparatives menées sur *Fusarium graminearum* et portant sur une large gamme de méthodes d'apprentissage statistique, allant de celles basées sur des scores linéaires aux réseaux de neurones profonds en passant par les forêts aléatoires, conduisent à des résultats peu satisfaisants.

L'objectif de notre travail est de proposer une nouvelle approche dans laquelle l'expression des gènes est vue comme un signal le long de la séquence de nucléotides support de la partie codante du gène. Ce nouveau paradigme pour l'apprentissage statistique *function-to-function*

de données fonctionnelles massives par d'autres données fonctionnelles tire profit de supports communs des signaux épigénomiques et d'expression pour valoriser des corrélations spatiales induites par des mécanismes connus de régulation de l'expression par le niveau d'accessibilité de la chromatine autour du codon d'initiation de chaque gène. La présentation démontre l'intérêt de cette approche à la fois en termes de performance de prédiction mais aussi pour mieux comprendre les relations entre épigénome et transcriptome.

**Mots-clés.** Apprentissage statistique, Données fonctionnelles massives, Intégration de données -omiques, Régression pour données fonctionnelles.

**Abstract.** Plant pathogens have to adapt to a variety of environmental stresses in order to grow and survive. Epigenetic variations can drive short-term phenotypic responses to those environmental stresses by modifications of the gene regulatory network, without any changes of the genomic sequence. This interplay between genomes, epigenomes and environment contributes to the emergence of new phenotypes, whose adaptability ensures more resilience. Whereas genetic mutations involved in species adaptation are intensively studied, epigenetic determinants of phenotypic innovation are so far poorly investigated, a major pitfall being the difficulty to adjust for the variations of the genetic background between generations. In this context, organisms with clonal reproduction are convenient models to study the specific contribution of epigenetics to adaptation mechanisms. The filamentous fungus *Fusarium graminearum* is an example of a highly resilient clonal reproduction plant pathogen producing a mycotoxin responsible for heavy damages in crops.

Variations of chromatin accessibility by changes of nucleosome positions are key epigenetic mechanisms regulating gene expression. High-throughput sequencing technologies (MAINE-seq) are especially designed to measure those whole-genome epigenetic variations. They generate complex and heterogeneous signals for which statistical genomics does not provide any standard data analysis routine so far. Statistical learning of the patterns of association between the epigenomic signal observed for each gene within a large genomic region covering the promoter and coding regions and gene expression measured using RNA-sequencing technologies (RNA-seq) offers some perspectives to understand epigenetic drivers of adaptability. However, comparative studies conducted on *Fusarium graminearum* and including a large panel of statistical learning methods, based on linear scores, aggregations of tree-based predictions or neural networks, show limited prediction performance.

We propose a new approach in which gene expression is viewed as a continuous signal over the coding region of a gene. This new *function-to-function* paradigm for statistical learning of massive functional responses by functional predicting variables takes advantage of common spatial supports for epigenomic and gene expression signals to leverage spatial correlations induced by proven transcriptional regulation mechanisms by chromatin accessibility in the neighborhood of the start codon of a gene. The extent to which this approach improves prediction performance is discussed in the presentation and prospective leads to better understand epigenetic regulation of transcription are deduced.

**Keywords.** Massive functional data, Statistical learning, Multi-omic data integration, Regression for functional data.

# 1 Introduction

Transcriptomic variations induced by many kinds of environmental stresses, especially heat stress, have been studied for a large variety of organisms. Identification of genes or more generally regulation pathways involved in those stress-induced variations generally results from so-called differential analyses consisting in whole-genome statistical tests for the comparison of mean expressions between contrasted experimental conditions. For example, such studies demonstrate that approximately 43% of genes in the filamentous fungus *Fusarium granimearum*, a clonal reproduction plant pathogen showing a high resilience to heat stresses (see Clairet *et al.*, 2023) have a modified gene expression level when exposed to a strong heat stress (15 minutes long exposition to 37 degrees Celsius). Moreover, complementary studies introducing a whole-genome mapping of epigenetic marks highlight correlations between gene expression levels and the presence or not of such marks on the genomic sequence of nucleotides within the neighborhood of those genes. Deciphering patterns of association between transcriptomic variations and epigenetic mechanisms involving modifications of the chromatin structure is yet a current challenge for bioinformatics and statistical genomics.

Availability of binding sites to transcription factors is notoriously affected by the positioning of nucleosomes along the sequence of nucleotide pairs in a broad region covering the promoter, coding and terminator regions of genes. How nucleosomes are positioned along the chromatin therefore plays a fundamental role in the regulation of transcription and subsequently gene expression. Nucleosome occupancy can be measured by MNase-assisted isolation of nucleosomes sequencing (MAINE-seq) technology, mapping regions of the genomes protected by nucleosomes. The former high-throughput technology generates signals providing a numeric measurement of chromatin accessibility at every base pair (bp) along the genome. In most studies involving such MAINE-seq data, raw signals of chromatin accessibility are reduced to summary statistics giving a general overview of the nucleosome occupancy over specific regions of interest in the genome. However, the dynamics of the transcription machinery suggests that numbers, amplitudes and positions of peaks and throughs in signals of chromatin accessibility should not be ignored, to account for a spatial correspondence between intervals of base pairs where the chromatin is accessible and some delayed transcriptional activity within the coding region of genes.

Identifying the specific patterns in curves of chromatin accessibility responsible for the regulation of gene expression is a key step to understand the epigenetic mechanisms driving phenotypes for improved adaptation of organisms to environmental stresses. The focus of our study (supported by the national research program Digit-Bio, INRAE, see https://digitbio-ia.github.io/) is on *F. granimearum*, whose epigenetic drivers are poorly known. Our goal is to investigate this question using statistical learning methods to establish association rules between patterns of MAINE-seq epigenomic data and RNA-seq measurements of gene expressions.

## 2 Data preparation and quality control

For each of the 14145 genes of *Fusarium granimearum*, three replicates of epigenomic signals are available on a region starting from 800 bp before the start codon to 800 bp after the stop codon. Those fixed values of 800 bp in the promoter and terminator regions of each gene has been chosen as a compromise between the necessity of leveraging epigenomic information outside of the coding region and the risk of interference with signals of too close genes, the genome of *Fusarium granimearum* being very dense. A data quality control procedure is first designed based on those three replicates, leading to the exclusion of 3350 genes, either showing inconsistencies across replicates, or with abnormal intervals of zeroes, or whose distance to other genes is smaller than 400 bp. Finally, a curve of log-transformed chromatin accessibility indices is obtained after averaging over the three replicates for the 10795 remaining genes. For each gene, three replicates of RNA-seq read counts within the coding region are also available. Gene expressions are log-transformed averages of those three replicates. The logarithm transformation applied on the raw chromatin accessibility indices and on the RNA-seq read counts is motivated by a strong over-dispersion of the data.

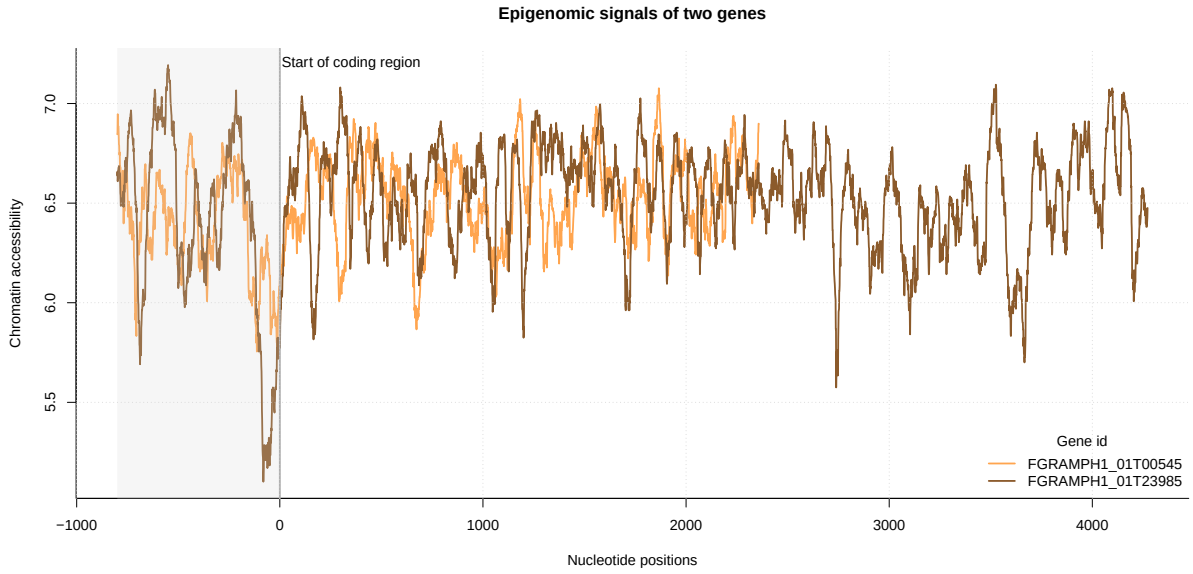



Figure 1: Epigenomic signals for two genes (`FGRAMPH1_01T00545` and `FGRAMPH1_01T23985`). The coding region of each gene starts at zero (plain vertical line). The promoter region is coloured in grey.

For illustration, Figure 1 displays the resulting epigenomic signals for two genes, whose identifiers are `FGRAMPH1_01T00545` and `FGRAMPH1_01T23985`, starting 800 bp before the start codon. First, note that the lengths of those genes and consequently of their epigenomic signals are different (1559 and 3473 bp), which raises a first major issue for designing statistical learning procedures based on those curves. More generally, gene lengths over the genome of

*Fusarium graminearum* are highly variable, as shown by the histogram displayed in Figure 2.

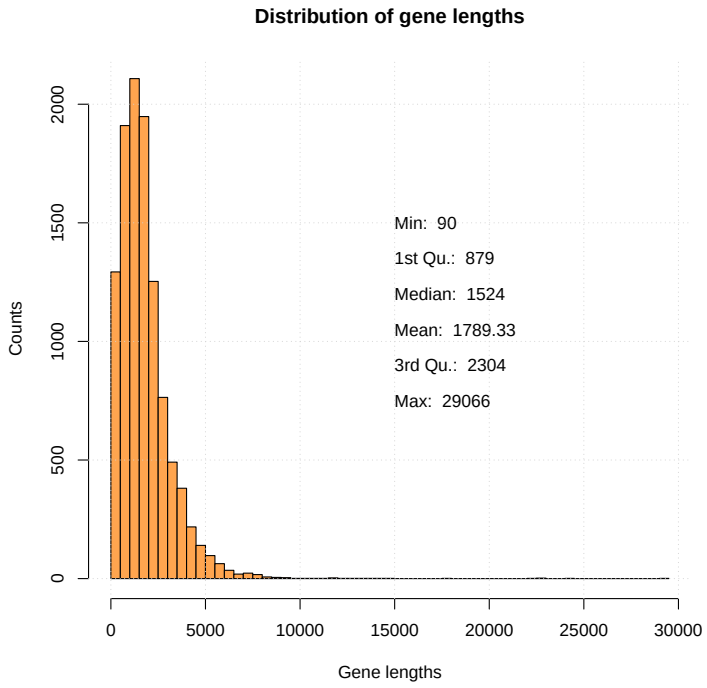



Figure 2: Histogram of gene lengths over the genome of *Fusarium graminearum*.

Many options have been considered to address this issue. First, reducing all signals to the same list of counts for histone modifications within bins of base pairs (see Singh *et al.*, 2016) or of geometrical summary statistics, such as latencies and amplitudes of peaks, or areas under the curve over some targeted regions, leads to a common definition of all explanatory variables for each gene. Unfortunately, how to draw this list of summary statistics is not guided by any biological consideration, which exposes to the risk of missing important patterns of association with gene expression. A second option preserving the potential prediction ability of whole epigenomic curves is to align all signals within the coding region, by taking a common grid of points for all genes, regularly spaced by the same fraction of the whole gene lengths. This option turns out to reduce one-to-one correlations between chromatin accessibility indices and gene expressions, raising suspicion about the relevance of such a transformation of epigenomic signals. The third option consists in restricting the study of epigenomic signals over an interval centered on the start codon, for example [-800 bp ; 800 bp] for all genes whose coding regions is at least longer than 800 bp. Indeed, as shown by the curve of correlations between chromatin accessibility indices and gene expressions displayed in Figure 3, the largest correlations are observed around the start codon of genes. More precisely, the negative peak of correlations at the end of the promoter region confirms that chromatin accessibility within this particular region favors the transcriptional activity.

In the presentation, we will focus on this last option, leading to a training dataset of joint

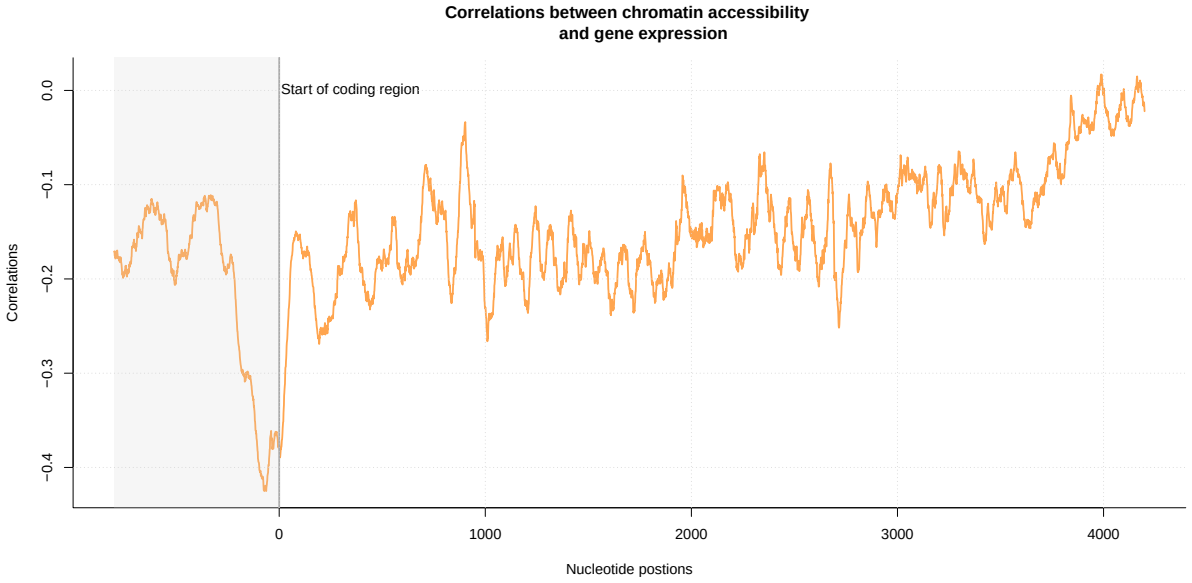



Figure 3: Correlations between chromatin accessibility indices at all nucleotide positions (limited to 4200 bp after the start codon, only 5% of genes being longer than 4200 bp) and gene expressions. The coding region of each gene starts at zero (plain vertical line). The promoter region is coloured in grey.

observations of the RNA-seq measurement of gene expression and discretized observations of the corresponding epigenomic signal on a grid of 1600 nucleotide positions around the start codon of each of 8440 genes. Note also that individuals in the present statistical learning issue are genes, which questions the usual independence assumption required to guarantee good properties of most standard estimation procedures. Indeed, gene expressions are notoriously driven by a regulatory network inducing a graph-structured stochastic dependence pattern across genes. This point will be discussed as prospective directions for improved prediction.

# 3 Statistical learning of gene expression based on epigenomic signals

In most prediction issues, the ideal objective in terms of prediction performance is to reach a cross-validated mean-squared error of prediction (MSEP) as close as possible to zero. In the present context, the ideal MSEP is unknown. It is indeed reasonable to consider that perfect prediction of gene expression from epigenomic signals exclusively would be a biological nonsense. Integration of additional -omic data, especially Hi-C data describing the three-dimensional conformation of chromatin, will be discussed as promising perspectives in the presentation.

A variety of statistical learning methods have been implemented to set up prediction

rules of gene expression from either raw epigenomic signals or reduced signals using B-spline coefficients in spline approximations with a large scope of dimensions for the basis:

- Penalized regression methods (Lasso, Ridge, Adaptive lasso and Elastic net with a fixed combination of the $\ell_1$ (0.9) and $\ell_2$ (0.1) norm of coefficients in the penalty term), where the penalty parameter minimizes a 10-fold cross-validated MSEP;
- Partial Least Squares (PLS) regression, where the number of PLS components minimizes a 10-fold cross-validated MSEP;
- Random forests (RF), where the number of variables to randomly sample at each split and the number of trees minimize the MSEP over a grid in a 10-fold cross-validation setup;
- Support Vector Regression (SVR);

The prediction performance of the state-of-the-art machine learning methods listed above and the CNN are reproduced in Table 1.

| | Ridge | Lasso | Adaptive Lasso | Elastic net | PLS | RF | SVR |
|---|---|---|---|---|---|---|---|
| RMSEP | 3.69 | 3.60 | 3.57 | 3.58 | 3.62 | 2.02 | 1.80 |

Table 1: Root Mean Squared Error of Prediction (RMSEP) of state-of-the-art machine learning methods for the prediction of gene expression using [-800 bb ; 800 bp] epigenomic signals. RMSEP is calculated in a 80% train-20% test cross-validation setup.

It turns out that the best prediction performance is reached by the random forests and the Support Vector Regression method, implemented either with the raw epigenomic signals or equivalently using a spline approximation with a large number of B-spline functions. Indeed, using a spline approximation does not improve prediction but speed up calculations. Yet, the former methods show a limited prediction performance.

# 4 The function-to-function prediction approach

One of the leads to improve the prediction performance of gene expression using epigenomic signals is motivated by the observation that the RNA-seq gene expression measurement is a total reads count over the coding region of genes, which can be viewed as a summary statistic masking a variety of dynamics of the transcriptional activity. For example, the two genes whose epigenomic signals are displayed in Figure 1 have indeed different epigenomic signals, especially at the end of the promoter region where the amplitude of the negative peaks identified as a possible marker of a large expression are different, but they have the same gene expression measurement. However, the locations within the coding region where the reads map exons of genes are distributed differently, as shown by their curves of cumulative

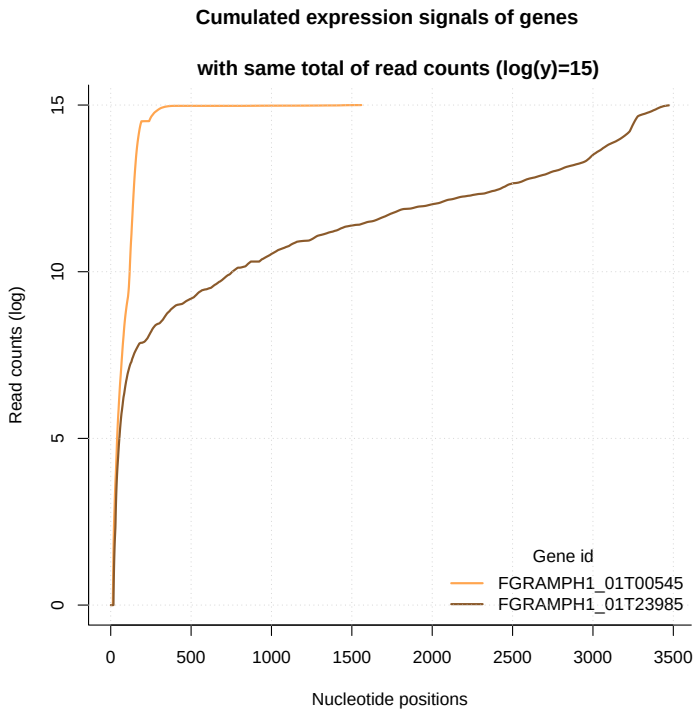



Figure 4: Cumulative numbers of reads within the coding region for the two genes whose epigenomic signals are displayed in Figure 1. The maximum value of each curve is reached before the stop codon and equals the total RNA-seq reads count.

log-transformed read counts, starting from the start codon up to the stop codon (see Figure 4).

In order to take advantage of a common spatial support of both epigenomic and transcriptional information, we propose to measure gene expression by the curve of cumulative reads count starting from the start codon and ending at 800 bp. Moreover, it turns out that the cumulative gene expression curve over this interval [1;800 bp] within the coding region shows an excellent ability to predict the total read counts. Indeed, Figure 5 demonstrates the accuracy of prediction of the total read counts by the cumulative gene expression curve over [1;800 bp], using PLS regression, with a 10-fold cross-validated correlation between observed and predicted value of 0.98.

The method we propose is two-step:

- Step 1: set up a prediction rule for the curve of cumulative gene expression in [1;800 bp] using epigenomic signals;
- Step 2: use PLS regression, whose performance is shown in Figure 5, to deduce the predicted total RNA-seq reads count.

The first step of the above method consists in predicting a curve, describing the cumulative

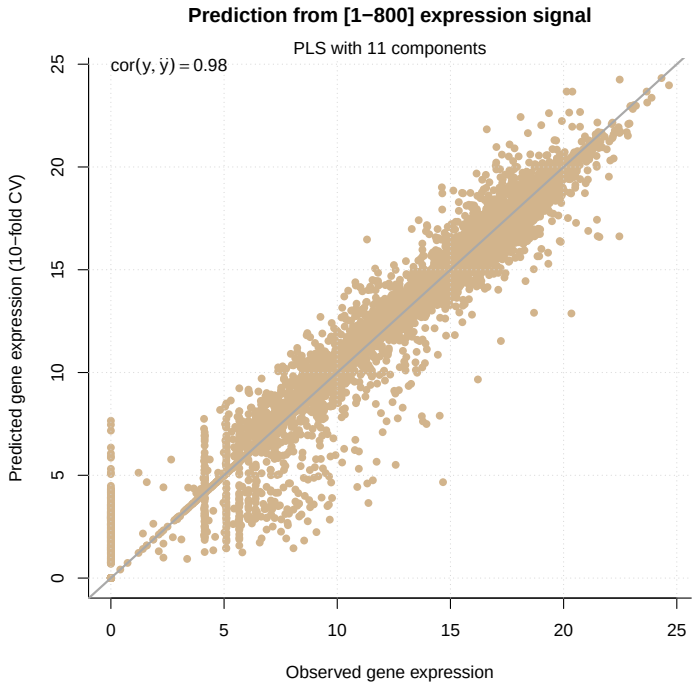



Figure 5: Performance of PLS regression in the prediction of the total RNA-seq reads count by the curve of cumulative gene expression in the interval [1;800 bp] starting at the start codon.

gene expression in [1;800 bp] by another curve, the epigenomic signal. For this *function-to-function* prediction task, two methods show good performance: random forests, where the loss function is the mean of squared differences between discretized observations and predictions of the curve of cumulative gene expressions, and a multi-head self-attention neural network. The added-value of using neural networks was previously mentioned in Singh *et al.* (2016), reporting a study in which a two-group expression level (high or low) is predicted by the counts for five pre-chosen histone modifications within adjacent 100 bp-bins around the start codon of each gene using Convolution Neural Networks (CNNs). Introducing self-attention layers in the neural network is inspired by the desirable properties of such mechanisms in language models, where the prediction of sequences by other sequences is similar to the present function-to-function prediction task.

The presentation will discuss the implementation of the above two-step method and show the improvements with respect to the direct prediction of the total RNA-seq reads count using the epigenomic signals. Promising perspectives for the current work will also be proposed, including integration of additional -omic data for a more complete description of the association between epigenome and transcriptome.

# Bibliography

Clairet, C., Lapalu, N., Simon, A., Soyer, J.L., Viaud, M., Zehraoui, E., Dalmais, B., Fudal, I. and Ponts, N. (2023) Nucleosome patterns in four plant pathogenic fungi with contrasted genome structures, *Peer Community Journal*, 3: e13.

Singh, R., Lanchantin, J., Robins, G., Qi, Y. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications, *Bioinformatics*, Volume 32, Issue 17, September 2016, Pages 639–648, https://doi.org/10.1093/bioinformatics/btw427

# Chaînes et processus de Markov

# MODÈLES AR FAIBLES MODULÉS PAR UNE CHAINE DE MARKOV CACHÉE

[1] Armel Bra, [2] Rabehasaina Landy & [3] Yacouba Boubacar Maïnassara

[1] kja.bra@univ-fcomte.fr

[2] landy.rabehasaina@univ-fcomte.fr

[3] Yacouba.BoubacarMainassara@uphf.fr

[1,2] Université Bourgogne-Franche-Comté,
Laboratoire de mathématiques de Besançon,
25030 Besançon, France.

[3] Université Polytechnique Hauts-de-France,
INSA Hauts-de-France,
CERAMATHS-Laboratoire de Matériaux,
Céramiques et de Mathématiques,
F-59313 Valenciennes, France.

**Résumé.** Dans ce document, nous présentons les propriétés asymptotiques de l'estimateur des moments pour les modèles autorégressifs (AR) intégrant des changements de régime markoviens où les erreurs sont non corrélées mais pas nécessairement indépendantes, avec l'hypothèse que les régimes ne sont pas directement observables. L'assouplissement des hypothèses concernant la non-indépendance des erreurs et la non-observabilité directe des régimes élargit significativement l'applicabilité de cette classe de modèles AR à changements de régimes. Nous donnons des conditions nécessaires pour prouver la consistance et la normalité asymptotique de l'estimateur des moments dans un cas particulier du modèle étudié. Une attention particulière est portée à l'estimation de la matrice de covariance asymptotique.

**Mots-clés.** Estimation, processus stationnaire, chaîne de Markov cachée, changement de régime, méthode des moments, consistance, normalité asymptomatique.

**Abstract.** In this document, we present the asymptotic properties of the moment estimator for autoregressive (AR) models incorporating Markov regime changes, where errors are uncorrelated but not necessarily independent, with the assumption that the regimes are not directly observable. Relaxing the assumptions regarding the non-independence of errors and the direct non-observability of regimes significantly broadens the applicability of this class of AR models with regime changes. We provide necessary conditions to prove the consistency and asymptotic normality of the moment estimator in a specific case of the model under study. Particular attention is paid to the estimation of the asymptotic covariance matrix.

**Keywords.** Estimation, stationary process, hidden Markov chain, regime switching, method of moments, consistency, asymptomatic normality.

# 1 Introduction

Depuis plusieurs décennies, l'analyse des séries temporelles est au centre de la recherche dans des domaines tels que l'économétrie et la finance. Les modèles ARMA, introduits par Box et Jenkins, constituent l'une des approches les plus classiques et les plus largement utilisées pour modéliser ces séries. Cependant, ces modèles conventionnels ont leurs limites, en particulier lorsqu'il s'agit de prendre en compte des changements soudains ou des transitions de régimes dans les données. (Voir par exemple Francq & Roussignol (1997)).

Dans ce document, notre attention est focalisée sur les modèles ARMA faibles modulés par une chaîne de Markov cachée. Lorsque nous parlons de modèles ARMA « faibles » , nous faisons référence à des modèles ARMA dans lesquels le bruit est non corrélé mais pas nécessairement indépendant. Le terme « caché » signifie que les états de la chaîne de Markov ne sont pas directement observables, mais qu'ils influencent néanmoins le comportement de la série temporelle. À l'inverse, nous qualifierons de modèles ARMA forts ceux pour lesquels le bruit est considéré comme indépendant et identiquement distribué (i.i.d). Ces nuances sont essentielles pour comprendre les spécificités et les défis de l'approche que nous abordons.

Malgré les complexités inhérentes à notre sujet d'étude, il est important de reconnaître les contributions significatives précédentes qui ont exploré des problématiques similaires sous des conditions plus restrictives. Nous pouvons citer entre autres Francq & Zakoıan (2001) et Francq & Gautier (2004) qui ont respectivement exploré les propriétés probabilistes et établi des conditions explicites assurant

la consistance et la normalité asymptotique, ainsi que l'estimation de la matrice de variance asymptotique de l'estimateur des moindres carrés, dans le cadre d'un modèle ARMA fort modulé par une chaîne de Markov observée. Poursuivant dans cette veine et sous certaines hypothèses essentielles de mélange et d'ergodicité, Boubacar Maïnassara & Rabehasaina (2020) ont étendu leurs résultats au cas des modèles ARMA faibles lorsque la chaîne de Markov est observée. Notre but est de compléter cette littérature déja riche en considérant des modèles ARMA faibles dont la chaîne de Markov est cachée.

Les applications d'un tel modèle sont vastes, de la détection des changements de régime dans les séries économiques à l'analyse des signaux biomédicaux, en passant par les prévisions financières. Les modèles ARMA faibles modulés par une chaîne de Markov cachée offrent une nouvelle perspective pour comprendre la dynamique sous-jacente des certaines séries temporelles.

# 2 Modèle et hypothèses

On dira qu'un processus $(X_t)_{t\in\mathbb{Z}}$ admet une représentation ARMA(p,q) faible modulé par une chaîne de Markov $(\delta_t)$ cachée si pour tout $t \in \mathbb{Z}$,

$$X_t = -\sum_{i=1}^{p} a_i^0(\delta_t) X_{t-i} + \sum_{j=1}^{q} b_j^0(\delta_t)\epsilon_{t-j} + \epsilon_t \tag{1}$$

où $\epsilon_t = f^0(\delta_t)\eta_t$ avec $(\eta_t)$ un processus stationnaire centré satisfaisant $\mathbb{E}(\eta_t\eta_{t'}) = \sigma^2 \mathbb{1}_{\{t=t'\}}$, $(\delta_t)$ une chaîne de Markov non observée à valeurs dans $\mathcal{E} := \{1, \cdots, K\}$, $f : \mathcal{E} \to \mathbb{R}$ et $(a^0(\delta_t))_{t\in\mathbb{Z}}, (b^0(\delta_t))_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$.

On désigne par $\theta_0$ le paramètre inconnu (les vrais paramètres). Formellement

$$\theta_0 := (a_i^0(s), b_i^0(s), p^0(k,k'), f^0(s); i = 1,\dots,p, j = 1,\dots,q; k, s \in \mathcal{E}; k' = 2,\dots,K) \in$$

$$\Theta := \Bigg\{ (a_i(s), b_j(s), p(k,k'), f(s); i = 1,\dots,p, j = 1,\dots,q, k, s \in \mathcal{E}, k' = 2,\dots,K) \in$$

$$\mathbb{R}^{(p+q)K} \times [0,1]^{K(K-1)} \times \mathbb{R}^K \text{ et } \forall\ k \in \mathcal{E}, \sum_{k'=2}^{K} p(k,k') \leq 1 \Bigg\}.$$

On pose

$$D_a := \begin{pmatrix} a_1^0 & & \\ & \ddots & \\ & & a_K^0 \end{pmatrix} \ , \ P' := \begin{pmatrix} p^0(1,1) & p^0(2,1) & \dots & p^0(K,1) \\ \vdots & \vdots & \vdots & \vdots \\ p^0(1,K-1) & p^0(2,K-1) & \dots & p^0(K,K-1) \\ 1-\sum_{j=1}^{K-1} p^0(1,j) & 1-\sum_{j=1}^{K-1} p^0(2,j) & \dots & 1-\sum_{j=1}^{K-1} p^0(K,j) \end{pmatrix}$$

et on considère les hypothèses suivantes :

- $(\mathbf{H}_1)$ Le processus $(X_t)$ est un processus ergodique.
- $(\mathbf{H}_2)$ Pour $\nu > 0$, le rayon spectral de $D_a^{4+2\nu} P'$ est strictement inférieur à 1.
- $(\mathbf{H}_3)$ Les processus $(\Delta_t)$ et $(\eta_t)$ sont indépendants.
- $(\mathbf{H}_4)$ $\sum_{h=0}^{\infty} \alpha_\eta(h)^{\frac{\nu}{2+\nu}} < \infty$ et $\mathbb{E}(|\eta_t|^{4+2\nu}) < \infty$ pour un certain $\nu > 0$.
- $(\mathbf{H}_5)$ L'intérieur $\mathring{\Theta}$ de $\Theta$ est non vide.
- $(\mathbf{H}_6)$ Il existe un sous-ensemble compact $\Theta^c$ de $\Theta$ contenant $\theta_0$.

Sous les hypothèses $(\mathbf{H}_2)$ et $(\mathbf{H}_3)$, les paramètres du modèle (1) lorsque $(p,q) = (0,1)$ satisfont pour tout $k \in \mathbb{Z}$,

$$c_{k,0} := \mathbb{E}\left(X_k X_0\right) = \sigma^2 \mathbf{1}'(D_a P')^k D_a^2 P'(I - D_a^2 P')^{-1} \pi_{f^2} := \psi_k(\theta_0), \qquad (2)$$

où $\mathbf{1}' = (1, \cdots, 1)$ est un vecteur de dimension $K$ et $\pi_{f^2} := \{f^2(1)\pi(1), \cdots, f^2(K)\pi(K)\}'$.

Un estimateur implicite de $\theta_0$ basé sur la méthode des moments est donné par $\hat{\theta}_n$ où $\hat{\theta}_n$ est le zéro de la fonction d'estimation $F^n(\theta) := \big(\psi_1(\theta) - \hat{c}_{1,0}, \cdots, \psi_{(p+q)K}(\theta) - \hat{c}_{(p+q)K,0}\big)$.

Une attention particulière sera accordée à la fonction $F^n$ afin de prouver les propriétés asymptotiques de $\hat{\theta}_n$.

# 3 Propriétés asymptotiques de l'estimateur des moments

Soient $\mathcal{F}_{-\infty}^t$ et $\mathcal{F}_{t+h}^{\infty}$ les $\sigma$-algèbres engendrés respectivement par $\{\eta_u : u \leq t\}$ et $\{\eta_u : u \geq t+h\}$. Afin de mesurer la dépendance temporelle du processus $(\eta_t)_{t\in\mathbb{Z}}$, nous introduisons les coefficients de mélange fort $(\alpha_\eta(h))_{h\in\mathbb{N}^\star}$ du processus stationnaire $(\eta_t)_{t\in\mathbb{Z}}$ comme suit

$$\alpha_\eta(h) := \sup_{A\in\mathcal{F}_{-\infty}^t, B\in\mathcal{F}_{t+h}^{\infty}} |P(A\cap B) - P(A)P(B)|.$$

**Théorème 1.** *Sous les hypothèses* $(\mathbf{H}_1)$, $(\mathbf{H}_2)$ *et* $(\mathbf{H}_3)$, *l'estimateur* $\hat{\theta}_n$ *converge presque sûrement vers* $\theta_0$.

**Théorème 2.** *Sous les hypothèses* $(\mathbf{H}_1)$ *à* $(\mathbf{H}_6)$, *nous avons*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, J^{-1}IJ^{-1}\right),$$

*avec* $J := J_{F^n}(\theta_0)$ *et* $I := I(\theta_0) = \lim_{n\to+\infty} \mathrm{Var}\left(\sqrt{n}F^n(\theta_0)\right) = \sum_{k=-\infty}^{\infty} \mathrm{Cov}(Y_t(\theta_0), Y_{t-k}(\theta_0))$,

*où* $J_{F^n}$ *désigne la matrice jacobienne de la fonction* $F^n$ *et* $Y_t(\theta_0) := \left(X_tX_{t+1}, \ldots, X_tX_{t+(p+q)K}\right)$.

**Remarque :** Bien que la fonction d'estimation $F^n$ dépende de $n$, il est important de noter que la matrice jacobienne $J_{F^n}$ n'est pas fonction de $n$. Ceci s'explique par le fait que, lors du calcul de la jacobienne de $F^n$, les termes en $\hat{c}_{k,0}$ sont éliminés, résultant en une expression de $J_{F^n}$ indépendante de $n$.

# 4 Estimation de la matrice de variance asymptomatique

Cette section vise à proposer un estimateur consistant pour la matrice de variance-covariance $\Omega := J^{-1}IJ^{-1}$ obtenue précédemment.Un estimateur simple pour la matrice $J$ dans le contexte de notre étude est donné par

$$\hat{J}_n := J_{F^n}(\hat{\theta}_n).$$

Pour estimer la matrice $I$, nous adopterons la méthode d'estimation paramétrique de la densité spectrale, telle qu'introduite par Berk (1974). Considérons $\hat{\varphi}_r(z)$ défini par $I_{(p+q)K} + \sum_{i=1}^{r} \hat{\varphi}_{r,i}z^i$, où les $\hat{\varphi}_{r,1}, \ldots, \hat{\varphi}_{r,r}$ représentent les coefficients obtenus par régression des moindres carrés de $\mathcal{Y}_t$ sur ses $r$ retards, soit $\mathcal{Y}_{t-1}, \ldots, \mathcal{Y}_{t-r}$ avec $\mathcal{Y}_t := Y_t(\theta_0) - \mathbb{E}[Y_t(\theta_0)]$. La variance empirique de ces résidus est notée $\hat{\Sigma}_{\hat{u}_r}$.

**Théorème 3.** *Supposons que les conditions du Théorème 2 soient satisfaites. De plus, admettons que* $\mathbb{E}|\eta_t|^{8+4\nu} < \infty$ *pour un certain* $\nu > 0$ , *que la matrice* $\Sigma_u$ *soit inversible, et que la suite* $(\mathcal{Y}_t)_{t\in\mathbb{Z}}$ *admet une représentation* $AR(\infty)$. *On suppose de plus que* $\|\varphi_i\| = o(i^{-2})$ *lorsque* $i \to \infty$, *et que les racines de* $\det(\boldsymbol{\varphi}) = 0$ *se situent en dehors du disque unité. Dans ces conditions, l'estimateur spectral pour la matrice* $\hat{I}^{SP}$, *défini par*

$$\hat{I}^{SP} := \hat{\boldsymbol{\varphi}}(1)^{-1}\hat{\Sigma}_{\hat{u}_r}\hat{\boldsymbol{\varphi}}(1)^{-1},$$

*converge en probabilité vers* $I$ *lorsque* $r = r(n) \to \infty$ *et* $r = o(n^{-1/3})$. *Par conséquent, un estimateur consistant de* $\Omega$ *est donné par*

$$\hat{\Omega} := \hat{J}_n^{-1}\hat{I}^{SP}\hat{J}_n^{-1}.$$

# Références

Berk, Kenneth N. 1974. Consistent autoregressive spectral estimates. *The Annals of Statistics*, 489–502.

Boubacar Maïnassara, Yacouba, & Rabehasaina, Landy. 2020. Estimation of weak ARMA models with regime changes. *Statistical Inference for Stochastic Processes*, **23**, 1–52.

Francq, Christian, & Gautier, Antony. 2004. Estimation of time-varying ARMA models with Markovian changes in regime. *Statistics & probability letters*, **70**(4), 243–251.

Francq, Christian, & Roussignol, Michel. 1997. On white noises driven by hidden Markov chains. *Journal of Time Series Analysis*, **18**(6), 553–578.

Francq, Christian, & Zakoıan, J-M. 2001. Stationarity of multivariate Markov–switching ARMA models. *Journal of Econometrics*, **102**(2), 339–364.

# Estimating the transitions of a Markov chain from incompletely observed paths in the presence of predictors

Daphné Aurouet Valentin Patilea

*Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France*
{*daphne.aurouet, valentin.patilea*}*@ensai.fr*

**Résumé.** Nous nous intéressons à l'estimation de la matrice de transition d'une chaîne de Markov à temps discret avec un nombre fini d'états en présence de covariables. Les données sont obtenues par la réalisation de trajectoires indépendantes de la chaîne, qui sont observées à des moments aléatoires. Cette étude est motivée par la circulation des billets de banque où une trajectoire correspond à un billet. En posant des hypothèses appropriées d'indépendance, nous construisons des estimateurs simples de la matrice de transition conditionnelle, sachant les valeurs des covariables, à l'aide de l'estimateur empirique et des racines $\ell$-ièmes des matrices de transition. L'effet de prédicteurs continus est pris en compte par un lissage à noyau et celui des covariables discrètes par une stratification. L'estimateur conditionnel de la matrice de transition peut être facilement appliqué à des jeux de données volimineux et transmis en continu, et celui-ci requiert peu de ressources informatiques. La performance de notre approche est illustrée sur des données simulées.

**Mots-clés.** Circulation des billets, Lissage à noyau, Chaînes de Markov, Racine $\ell-$ième d'une matrice stochastique

**Abstract.** We are interested in estimating the transition matrix of a discrete time, finite Markov chain in the presence of covariates. The data points are obtained from independent sample paths of the chain which are observed at random times. This study was motivated by the circulation of banknotes where a sample path corresponds to a banknote. Based on suitable independence assumptions, we build simple conditional transition matrix estimates, given the covariate vector value, using empirical estimators of transition probabilities and $\ell-$th roots of the transition matrices. The effect of continuous covariates is accounted for by kernel smoothing and the one of discrete covariates via stratification. The conditional transition matrix estimator can be easily applied to large streaming data and requires low computer resources. The performance of our approach is illustrated using simulated data.

**Keywords.** Banknote circulation, Kernel smoothing, Markov chains, $\ell-$root of a stochastic matrix

## 1 Introduction

In this work, we consider discrete-time Markov chains with finite state space. Such models are commonly found in real-world applications such as statistical mechanics (Seneta, 2016),

finance (Israel et al., 2001), predictive maintenance (Tamaloussi and Bouzaouit, 2020) *etc.* Their appeal comes from their memory-less property, ease of interpretation, and simple computations. The systems under consideration in this study are homogeneous, irreducible, and aperiodic, but are observed at random times only. The last characteristic has been little studied in theory, although it is commonly encountered in practice.

Observing the state of a Markovian system at every unit of time, as is usually assumed for Markov chains, can be costly or impossible. We therefore take a more pragmatic approach and allow observations to occur at random discrete times. Consequently, some jumps between successive observations may occur and go unobserved. We also assume that the system is influenced by additional external factors (or covariates), gathered in a vector with continuous or discrete components that do not change over time. Conditionally on the value of the covariate vector, the Markov chain is homogeneous. Our objective is thus to estimate the conditional transition matrix from incompletely observed sample paths of a discrete-time Markov chain given the values of the covariate vector. The lengths of the sample paths of the chain are bounded, but for the applications we have in mind, we can consider that a large number of independent sample paths are generated. While this setting looks similar to discrete-time semi-Markov chains, it adds the difficulty of unobserved transitions.

A related problem has been studied by Barsotti et al. (2014). The authors considered a method to build an estimator of the (unconditional) transition matrix under the assumption that it has some zero entries. Here, in the presence of covariates, we propose a kernel smoothing estimator that remains simple and flexible. Our method is constructed at the cost of more restrictive assumptions on the observation mechanism. However, our assumptions appear to be well suited to the study of the banknote circulation. We leverage the homogeneity of our process given the covariate, and consider the $\ell-$power of the conditional transition matrix. The elements of this matrix can be simply estimated by smooth empirical probability estimators, yielding a matrix estimate which is a stochastic matrix. Under certain conditions, we can then compute the $\ell-$root of the estimated $\ell-$power of the conditional transition matrix. The existence and stochasticity of the $\ell-$root is a challenging aspect of the proposed approach, which will be pointed out below. It is related to a well known, and still open problem in probability theory and matrix algebra. Several positive integer values $\ell$ can be considered, and we propose to aggregate the resulting conditional transition matrix estimators by a weighted average.

Section 2 details the model framework and necessary assumptions, and Section 3 introduces the proposed estimator. Section 4 describes the performance of a finite sample.

# 2 Markov Chains with randomly observed paths

Let $\mathbf{X} = (X_t)_{t\geq 1}$ by a Markov chain with values in the finite set $\mathcal{S} = \{1, 2, \ldots, S\}$. Let $\mathbf{Z}$ be a random vector of covariates (predictors) which for simplicity we suppose independent of the time $t$.

## 2.1 The model

The covariate vector can be decomposed into a sub-vector of continuous predictors $\mathbf{Z}^c \in \mathbb{R}^{\boldsymbol{d}_c}$ and a sub-vector of discrete ones $\mathbf{Z}^d \in \mathbb{R}^{\boldsymbol{d}_d}$. In the following, $\mathbf{z} = (\mathbf{z}^c, \mathbf{z}^d) \in \mathbb{R}^{\boldsymbol{d}_c} \times \mathbb{R}^{\boldsymbol{d}_d}$ is any value in the support of $\mathbf{Z}$. We impose the following assumption on the conditional distribution of $\mathbf{X}$ given the covariate vector value.

(H1) Conditionally given $\mathbf{Z} = \mathbf{z}$, the discrete time process $\mathbf{X}$ is a homogeneous, irreducible and aperiodic Markov chain with transition probabilities

$$p_{ij}(\mathbf{z}) = \mathbb{P}(X_{t+1} = j \mid X_t = i, \mathbf{Z} = \mathbf{z}), \qquad i, j \in \mathcal{S}.$$

Let $\boldsymbol{P}(\mathbf{z})$ be the conditional transition matrix $\mathbf{X}$ given $\mathbf{Z} = \mathbf{z}$.

For the applications we have in mind, we observe several independent paths of $\mathbf{X} = (X_t)_{t\geq 1}$, and for each path we also observe a random draw of $\mathbf{Z}$. Unfortunately, the independent paths of $\mathbf{X}$ are incompletely observed. More precisely, instead of being observed at each integer $t \geq 1$, a sample paths is observed at the integer random times $1 \leq T_0 < T_1 < \ldots < T_k < T_{k+1} < \ldots \leq t_{\max}$. The non-random integer $t_{\max}$ defines the observation window for the transitions of a sample paths, and this value is given. Let

$$\tau_k = T_k - T_{k-1} \qquad \text{and} \qquad Y_k = X_{T_k}, \qquad k \geq 1.$$

The observed sample consists of independent realizations of $\mathbf{Y} = (Y_k, \tau_k, \mathbf{Z})_{k\geq 1}$, observed as long as $\tau_1 + \cdots + \tau_k \leq t_{\max}$. Therefore, we can impose $\tau_k \in \{1, 2, \ldots, L\}$ for some integer $L \geq 1$. Some additional assumptions on the distribution of $\mathbf{Y}$ are considered.

(H2) For any $i, j \in \mathcal{S}$, and $\ell, \ell' \in \{1, 2, \ldots, L\}$,

$$\begin{aligned}\mathbb{P}(Y_{k+1} = j, \tau_{k+1} = \ell \mid Y_k = i, \tau_k = \ell', \mathbf{Z} = \mathbf{z}, Y_{k-1}, \tau_{k-1}, \ldots, Y_1, \tau_1, T_0) \\ = \mathbb{P}(Y_{k+1} = j, \tau_{k+1} = \ell \mid Y_k = i, \mathbf{Z} = \mathbf{z}), \qquad \forall k \geq 1. \quad (1)\end{aligned}$$

(H3) For any $i, j \in \mathcal{S}$ and $t' > t \geq 1$,

$$\mathbb{P}(X_{T_{k+1}} = j \mid X_{T_k} = i, T_{k+1} = t', T_k = t, \mathbf{Z} = \mathbf{z}) = \left(\boldsymbol{P}^{t'-t}(\mathbf{z})\right)_{ij}, \qquad \forall k \geq 1. \quad (2)$$

Assumption (1) imposes a lack-of-memory condition for $(Y_k, \tau_k)_{k\geq 1}$, given $\mathbf{Z} = \mathbf{z}$. In particular, it implies that conditionally given $\mathbf{Z} = \mathbf{z}$, the process $(Y_k, \tau_k)_{k\geq 1}$ is a homogeneous Markov chain with a finite state space. Condition (2) can hold true even when the random times $(T_k)$ are not independent of $\mathbf{X}$.

Using (1) and (2) with $\ell = t' - t$, we get,

$$\begin{aligned}\mathbb{P}(Y_{k+1} = j \mid Y_k = i, \tau_{k+1} = \ell, \mathbf{Z} = \mathbf{z}) &= \mathbb{P}(Y_{k+1} = j \mid Y_k = i, \tau_{k+1} = \ell, T_k, \mathbf{Z} = \mathbf{z}) \\ &= \mathbb{P}(X_{T_k+\ell} = j \mid X_{T_k} = i, T_{k+1} = T_k + \ell, T_k, \mathbf{Z} = \mathbf{z}) = \left(\boldsymbol{P}^{\ell}(\mathbf{z})\right)_{ij}.\end{aligned}$$

Finally, we obtain

$$\mathbb{P}(Y_{k+1}=j,\tau_{k+1}=\ell \mid Y_k=i,\mathbf{Z}=\mathbf{z}) = \left(\boldsymbol{P}^{\ell}(\mathbf{z})\right)_{ij} \mathbb{P}(\tau_{k+1}=\ell \mid Y_k=i,\mathbf{Z}=\mathbf{z}). \tag{3}$$

From (3), we rewrite for $i,j \in \mathcal{S}$, $\mathbf{z}=(\mathbf{z}_c,\mathbf{z}_d)$ and $1 \leq \ell \leq L$,

$$\boldsymbol{P}^{\ell}(\mathbf{z}) = \boldsymbol{A}_{\ell}(\mathbf{z}) \quad \text{where} \quad (\boldsymbol{A}_{\ell}(\mathbf{z}))_{ij} = \frac{\mathbb{P}(Y_{k+1}=j,\tau_{k+1}=\ell \mid Y_k=i,\mathbf{Z}=\mathbf{z})}{\mathbb{P}(\tau_{k+1}=\ell \mid Y_k=i,\mathbf{Z}=\mathbf{z})}.$$

The matrix $\boldsymbol{A}_{\ell}(\mathbf{z})$ is thus the $\ell-$th power of the conditional transition matrix we want to estimate. It is worth noting that $\boldsymbol{A}_{\ell}(\mathbf{z})$ is a stochastic matrix, that is a square matrix with non-negative elements with each row summing to 1. Moreover, only observable variables are used to define the elements of $\boldsymbol{A}_{\ell}(\mathbf{z})$. A natural idea is then to estimate $\boldsymbol{A}_{\ell}(\mathbf{z})$ and next define an estimator of $\boldsymbol{P}(\mathbf{z})$ as the $\ell-$roots of $\boldsymbol{A}_{\ell}(\mathbf{z})$, provided that such matrix root exists and is a stochastic matrix.

## 2.2 The transition matrix and the $\ell-$roots of stochastic matrices

For any squared matrix $\boldsymbol{Q}$, the series $\sum_{k=0}^{\infty} \boldsymbol{Q}^k/k!$, converges with respect to any matrix norm. The limit is denoted $\exp(\boldsymbol{Q})$ and defines the exponential of $\boldsymbol{Q}$.

Let $\mathbf{z}$ be fixed. If the matrix $\boldsymbol{Q}$ is such that $\boldsymbol{P}(\mathbf{z}) = \exp(\boldsymbol{Q})$, then $\boldsymbol{P}^{\ell}(\mathbf{z}) = \exp(\ell\boldsymbol{Q}) = \boldsymbol{A}_{\ell}(\mathbf{z})$, for all $\ell \geq 1$. Moreover, by definition $\ell\boldsymbol{Q}$ is a logarithm of $\boldsymbol{A}_{\ell}(\mathbf{z})$, denoted $\log(\boldsymbol{A}_{\ell}(\mathbf{z}))$. Conversely, we can write

$$\boldsymbol{P}(\mathbf{z}) = \exp(\{\ell\boldsymbol{Q}\}/\ell) = \exp(\log(\boldsymbol{A}_{\ell}(\mathbf{z}))/\ell). \tag{4}$$

See, for example, Norris (1998) and Higham (2008) for the formal definitions of the exponential and logarithm of a matrix. The existence of the matrix $\boldsymbol{Q}$ is a classical but not completely solved problem in matrix algebra and numerical analysis, related to the so-called embedding problem in Markov chains theory. See Kingman (1962).

To discuss the existence of the representation (4) for the conditional transition matrix, we can start by searching conditions under which $\log(\boldsymbol{A}_{\ell}(\mathbf{z}))$ is well defined. Let $\boldsymbol{A}$ be a real matrix. The following properties are stated in (Higham, 2008, Theorem 1.31 and 11.2).

- If $\boldsymbol{A}$ has no eigenvalues in $(-\infty,0]$, there exists a unique logarithm of $A$ which is real. This matrix, say $\boldsymbol{X}$, is referred to as the *principal logarithm of* $\boldsymbol{A}$ and $\exp(\boldsymbol{X}) = \boldsymbol{A}$.
- If $\boldsymbol{A}$ has no eigenvalues in $(-\infty,0]$, then for any $\alpha \in [-1,1]$ we have $\log(\boldsymbol{A}^{\alpha}) = \alpha\log(\boldsymbol{A})$.

Since the exponential of a squared matrix always converge, the remaining and most difficult problem is to guarantee that $\exp(\log(\boldsymbol{A})/\ell)$ is a stochastic matrix. Several sufficient conditions have been provided, see for example Cuthbert (1972, 1973), Higham and Lin (2011) and the references therein.

In addition to the problem of the validity of the representation (4) for the theoretical matrix $\boldsymbol{P}(\mathbf{z})$, there is also the question whether $\exp(\log(\boldsymbol{A}_\ell(\mathbf{z}))/\ell)$ remains a stochastic matrix when $\boldsymbol{A}_\ell(\mathbf{z})$ is replaced by an estimator. Israel et al. (2001) and Bladt and Sørensen (2005) address this aspect and provide some partial answers and remedies.

The problem becomes even more complex in our context in the presence of covariates. For some values $\mathbf{z}$ the matrix $\exp(\log(\boldsymbol{A}_\ell(\mathbf{z}))/\ell)$ may be a stochastic matrix when $\boldsymbol{A}_\ell(\mathbf{z})$ is replaced by an estimate, but this may not happen for other values $\mathbf{z}$.

# 3 Estimating the conditional transition matrix

The sample consists of the data points corresponding to $N$ independent sample paths of $(Y_k, \tau_k)_{k\geq 1}$ and $N$ independent realizations of the covariate vector $\mathbf{Z}$. More precisely, for each $1 \leq m \leq N$, the vectors $(Y_{m,k}, \tau_{m,k}, \mathbf{Z}_m)$, $1 \leq k \leq M_m$, are observed. The integers $M_m$ are bounded by $L$.

Let us fix $1 \leq \ell \leq L$. For any $\mathbf{z} = (\mathbf{z}^c, \mathbf{z}^d)$ in the support of $\mathbf{Z}$, we define the estimator of the element $(i, j)$ of the matrix $\boldsymbol{A}_\ell(\mathbf{z})$ as

$$\left(\widehat{\boldsymbol{A}}_\ell(\mathbf{z})\right)_{ij} = \frac{\sum_{m=1}^{N}\sum_{k=1}^{M_m} \mathbb{1}\left\{Y_{m,k+1} = j, \tau_{m,k+1} = \ell, Y_{m,k} = i, \mathbf{Z}_m^d = \mathbf{z}^d\right\} \boldsymbol{K}_{\boldsymbol{h}}\left(\mathbf{Z}_m^c - \mathbf{z}^c\right)}{\sum_{m=1}^{N}\sum_{k=1}^{M_m} \mathbb{1}\left\{\tau_{m,k+1} = \ell, Y_{m,k} = i, \mathbf{Z}_m^d = \mathbf{z}^d\right\} \boldsymbol{K}_{\boldsymbol{h}}\left(\mathbf{Z}_m^c - \mathbf{z}^c\right)}. \quad (5)$$

The rule $0/0 = 0$ applies. Here, for any $\boldsymbol{u} \in \mathbb{R}^{\boldsymbol{d}_c}$ with components $u^{(1)}, \ldots, u^{(\boldsymbol{d}_c)}$, we define

$$\boldsymbol{K}_{\boldsymbol{h}}(\boldsymbol{u}) = \prod_{l=1}^{\boldsymbol{d}_c} K(u^{(l)}/h^{(l)}),$$

with $\boldsymbol{h} = (h^{(1)}, \ldots, h^{(\boldsymbol{d}_c)})$ a vector of bandwidths and $K(\cdot)$ an univariate kernel, for instance the standard Gaussian density. The components of the bandwidth vector $\boldsymbol{h}$ are defined as

$$h^{(l)} = \{M_1 + \cdots + M_N\}^{-\alpha}\, \widehat{\sigma}_l, \qquad l = 1, \cdots, \boldsymbol{d}_c, \quad \text{with} \quad \alpha = 1/(4 + \boldsymbol{d}_c), \quad (6)$$

and $\widehat{\sigma}_l$ the empirical standard deviation of the $l-$th component of the sub-vector $\mathbf{Z}^c$ gathering the continuous components of the covariate vector.

Assume that $\widehat{\boldsymbol{A}}_\ell(\mathbf{z})$ has no eigenvalues in $(-\infty, 0]$, and there exists a unique logarithm of $\widehat{\boldsymbol{A}}_\ell(\mathbf{z})$ which is a matrix with real elements. Then we can define an estimator of $\widehat{\boldsymbol{P}}(\mathbf{z})$ as the $\ell-$root of $\widehat{\boldsymbol{A}}_\ell(\mathbf{z})$ computed as

$$\widehat{\boldsymbol{P}}_\ell(\mathbf{z}) = \exp\left\{\frac{1}{\ell}\log\left(\widehat{\boldsymbol{A}}_\ell(\mathbf{z})\right)\right\}.$$

Whenever this estimator $\widehat{\boldsymbol{P}}(\mathbf{z})$ exists and is a stochastic matrix, we get an estimator of the conditional transition matrix $\boldsymbol{P}(\mathbf{z})$ given $\mathbf{Z} = \mathbf{z}$ based on the transitions observed after $\ell$ periods of time. In order to better exploit the information carried by the sample, we can

consider several $\ell$ values, typically those with the largest frequencies. Let $[\underline{L}, \overline{L}]$ a range of $\ell$ and let

$$\widehat{\boldsymbol{P}}(\mathbf{z}) = \frac{1}{\sum_{\ell=\underline{L}}^{\overline{L}} \widehat{\pi}_l} \sum_{\ell=\underline{L}}^{\overline{L}} \widehat{\pi}_l \widehat{\boldsymbol{P}}_\ell(\mathbf{z}),$$

be an aggregated estimator of the conditional transition matrix $\boldsymbol{P}(\mathbf{z})$ given $\mathbf{Z} = \mathbf{z}$. Here, $\widehat{\pi}_l$ is the empirical estimator of the marginal probability $\mathbb{P}(\tau_k = \ell)$.

It is worth noting that our estimation approach allows for very large datasets, in particular for streaming data. More precisely, few more observations can be added on each sample path, and, more important, many other sample paths can observed. The numerator and the denominator in (5) can be easily updated in such situations, without low memory resources and computational complexity. The bandwidth rule (6) guarantees the necessary gradual decrease with the new observations.

The asymptotic behavior of our estimators $\widehat{\boldsymbol{P}}_\ell(\mathbf{z})$, and the aggregated version $\widehat{\boldsymbol{P}}(\mathbf{z})$ can be derived from the asymptotic behavior of the estimators $\widehat{\boldsymbol{A}}_\ell(\mathbf{z})$. The asymptotic of these latter estimators can be studied by the standard tools used for the kernel regression with discrete responses. Theoretical grounds for the recursive versions of $\widehat{\boldsymbol{A}}_\ell(\mathbf{z})$, which are better suited with streaming data, can be obtained from the existing theoretical results on recursive kernel regression.

# 4 Empirical evidence

To construct our simulation design, we first consider a pilot stochastic matrix $\boldsymbol{B}$ for which the $\ell-$roots exist and are stochastic. The conditional transition matrix $\boldsymbol{P}(\mathbf{z})$ given $\mathbf{Z} = \mathbf{z}$ is then constructed by perturbing the pilot matrix $\boldsymbol{B}$, with the perturbation depending on $\mathbf{z}$. The covariate vector $\mathbf{Z}$ has up to three components. The discrete component of $\mathbf{Z}$ is a Bernoulli variable with parameter $p = 0.7$, while the continuous components are generated using a Beta distribution.

Independently for each $1 \leq m \leq N$, we draw the initial value $Y_{m,1}$ from a discrete uniform distribution over $\mathcal{S}$. Next, given $Y_{m,k}$, the $\tau_{m,k+1}$ is obtained as a random draw from a Poisson distribution with parameter $\lambda$ to which we add the value 1. The parameter $\lambda$ depends on the value : it is equal to 10 if $Y_{m,k} \in \{1, 2\}$ and equal to 15 if $Y_{m,k} \in \{3, \cdots, S\}$. Finally, given $\mathbf{Z}^{(m)} = \mathbf{z}$, $\tau_{m,\ell} = \ell$ and $Y_{m,k} = i$, we draw $Y_{m,k+1}$ from a multinomial distribution with the vector of parameters equal to the $i-$th row of the matrix $\boldsymbol{P}^\ell(\mathbf{z})$. The procedure is repeated $M_m$ steps where $M_m$ is the smallest integer such that $\sum_{k=1}^{M_m} \tau_{m,k} \geq t_{\max}$. We set $t_{\max} = 20$ in our experiences. The performance of the estimators of the conditional transition matrices is evaluated with the spectral norm of the error $\widehat{\boldsymbol{P}}(\mathbf{z}) - \boldsymbol{P}(\mathbf{z})$ (for any matrix $\boldsymbol{U}$, the square of the spectral norm is defined by $\|\boldsymbol{U}\|_2^2 = \lambda_{\max}\left(\boldsymbol{U}^\top \boldsymbol{U}\right)$, where for any positive semi-definite matrix $\boldsymbol{C}$, $\lambda_{\max}(\boldsymbol{C})$ is the largest eigenvalue of $\boldsymbol{C}$).

Table 1 and Figure 1 illustrate the performance of the estimator over 200 replications of experiments on different sample and state space sizes and different values of the predictor vector. It reveals the convergence of our estimator towards the true transition matrix for

most of the different scenarios tested. In very few cases, for some covariate vector values $\mathbf{z}$, the consistency does not seem guaranteed, most likely because of the $\ell-$roots of the empirical estimates of the matrices $\widehat{\boldsymbol{A}}_\ell(\mathbf{z})$ do not exist. See also the discussion in the section 2.2 above.

For larger state spaces, the number of parameters to be estimated is larger, and hence the performance is poorer. It is worth noting that a better accuracy for the estimators is achieved when the diagonal of the matrix $\boldsymbol{P}(\mathbf{z})$ is to a large extent dominant.

Figure 1: Median of the spectral norm of the errors for the multivariate predictor case, over different sample sizes from $N = 200$ to $N = 32400$, for a Markov chain with $S = 3$ states. The vector $\mathbf{Z}$ has a binary component $z_d$ and two continuous components $z_{c,1}$ and $z_{c,2}$. The value $t_{\max}$ is set equal to 20. Results obtained from $R = 200$ replications.

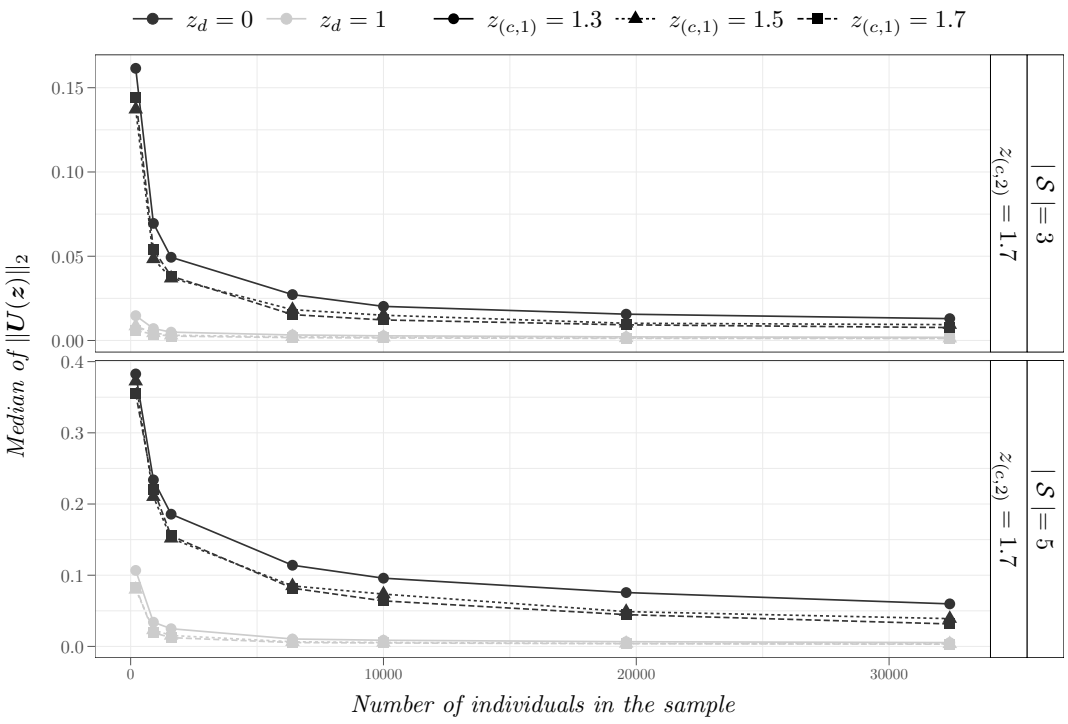


Table 1: Median of the spectral norm of the errors for the multivariate predictor case, over different sample sizes, for a Markov chain with $S \in \{3, 5, 9\}$ states. The cases of no predictor, one predictor, and three predictors (where $\mathbf{Z}$ has two continuous components and a binary component $z_d$) are considered. The value $t_{\max}$ is set equal to 20. Results obtained from $R = 200$ replications.

| | | | $N$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Predictors | $S$ | **z** | **200** | **900** | **1600** | **6400** | **10000** | **19600** | **32400** |
| No predictor | **3** | | 0.0675 | 0.0209 | 0.0135 | 0.0060 | 0.0048 | 0.0033 | 0.0023 |
| | **5** | | 0.1830 | 0.0800 | 0.0549 | 0.0184 | 0.0132 | 0.0086 | 0.0062 |
| | **9** | | 0.4915 | 0.2025 | 0.1560 | 0.0784 | 0.0636 | 0.0433 | 0.0330 |
| One predictor | **3** | (1.3) | 0.1096 | 0.0755 | 0.0702 | 0.0493 | 0.0469 | 0.0396 | 0.0333 |
| | | (1.5) | 0.0793 | 0.0491 | 0.0442 | 0.0336 | 0.0331 | 0.0262 | 0.0209 |
| | | (1.7) | 0.0808 | 0.0518 | 0.0413 | 0.0270 | 0.0218 | 0.0158 | 0.0126 |
| | **5** | (1.3) | 0.2053 | 0.1790 | 0.1696 | 0.1533 | 0.1499 | 0.1397 | 0.1328 |
| | | (1.5) | 0.1648 | 0.1025 | 0.0958 | 0.0836 | 0.0779 | 0.0715 | 0.0690 |
| | | (1.7) | 0.2184 | 0.1117 | 0.0969 | 0.0745 | 0.0698 | 0.0692 | 0.0594 |
| | **9** | (1.3) | 0.4493 | 0.3414 | 0.3468 | 0.3408 | 0.3268 | 0.3325 | 0.3246 |
| | | (1.5) | 0.3971 | 0.2239 | 0.1980 | 0.1876 | 0.1876 | 0.1751 | 0.1764 |
| | | (1.7) | 0.4785 | 0.1699 | 0.1408 | 0.1193 | 0.1157 | 0.1093 | 0.1132 |
| Three predictor | **3** | (1.3, 1.3, 0) | 0.1754 | 0.0730 | 0.0532 | 0.0325 | 0.0265 | 0.0199 | 0.0174 |
| | | (1.3, 1.3, 1) | 0.0205 | 0.0096 | 0.0077 | 0.0047 | 0.0041 | 0.0034 | 0.0029 |
| | | (1.3, 2.9, 0) | 0.1467 | 0.0699 | 0.0446 | 0.0131 | 0.0112 | 0.0078 | 0.0065 |
| | | (1.3, 2.9, 1) | 0.0020 | 0.0005 | 0.0005 | 0.0005 | 0.0005 | 0.0005 | 0.0005 |
| | | (1.7, 1.3, 0) | 0.1452 | 0.0568 | 0.0427 | 0.0207 | 0.0148 | 0.0118 | 0.0102 |
| | | (1.7, 1.3, 1) | 0.0083 | 0.0053 | 0.0043 | 0.0029 | 0.0022 | 0.0020 | 0.0017 |
| | | (1.7, 1.7, 0) | 0.1440 | 0.0540 | 0.0382 | 0.0154 | 0.0122 | 0.0093 | 0.0077 |
| | | (1.7, 1.7, 1) | 0.0059 | 0.0037 | 0.0026 | 0.0017 | 0.0015 | 0.0012 | 0.0011 |
| | | (1.7, 2.9, 0) | 0.1258 | 0.0491 | 0.0183 | 0.0072 | 0.0055 | 0.0046 | 0.0040 |
| | | (1.7, 2.9, 1) | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 |
| | **5** | (1.3, 1.3, 0) | 0.3478 | 0.2313 | 0.1890 | 0.1204 | 0.1088 | 0.0940 | 0.0870 |
| | | (1.3, 1.3, 1) | 0.1128 | 0.0510 | 0.0391 | 0.0156 | 0.0126 | 0.0091 | 0.0080 |
| | | (1.3, 1.7, 0) | 0.3827 | 0.2339 | 0.1857 | 0.1140 | 0.0958 | 0.0757 | 0.0597 |
| | | (1.3, 1.7, 1) | 0.1067 | 0.0340 | 0.0248 | 0.0104 | 0.0086 | 0.0066 | 0.0054 |
| | | (1.3, 2.9, 0) | 0.3526 | 0.2329 | 0.1725 | 0.0782 | 0.0614 | 0.0359 | 0.0268 |
| | | (1.3, 2.9, 1) | 0.0817 | 0.0128 | 0.0080 | 0.0012 | 0.0012 | 0.0011 | 0.0013 |
| | | (1.7, 1.3, 0) | 0.3412 | 0.2464 | 0.1770 | 0.0978 | 0.0809 | 0.0582 | 0.0448 |
| | | (1.7, 1.3, 1) | 0.0896 | 0.0316 | 0.0187 | 0.0081 | 0.0070 | 0.0055 | 0.0044 |
| | | (1.7, 1.7, 0) | 0.3551 | 0.2209 | 0.1553 | 0.0818 | 0.0639 | 0.0446 | 0.0315 |
| | | (1.7, 1.7, 1) | 0.0834 | 0.0193 | 0.0123 | 0.0054 | 0.0048 | 0.0037 | 0.0031 |
| | | (1.7, 2.9, 0) | 0.3265 | 0.1804 | 0.1335 | 0.0607 | 0.0387 | 0.0210 | 0.0132 |
| | | (1.7, 2.9, 1) | 0.0676 | 0.0085 | 0.0009 | 0.0005 | 0.0005 | 0.0005 | 0.0005 |

# References


Barsotti, F., De Castro, Y., Espinasse, T., and Rochet, P. (2014). Estimating the transition matrix of a Markov chain observed at random times. *Statistics & Probability Letters*, 94:98–105.

Bladt, M. and Sørensen, M. (2005). Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):395–410.

Cuthbert, J. R. (1972). On uniqueness of the logarithm for Markov semi-groups. *J. London Math. Soc. (2)*, 4:623–630.

Cuthbert, J. R. (1973). The logarithm function of finite-state Markov semi-groups. *J. London Math. Soc. (2)*, 6:524–532.

Higham, N. J. (2008). *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Theory and computation.

Higham, N. J. and Lin, L. (2011). On $p-$th roots of stochastic matrices. *Linear Algebra and its Applications*, 435(3):448–463.

Israel, R. B., Rosenthal, J. S., and Wei, J. Z. (2001). Finding generators for markov chains via empirical transition matrices, with applications to credit ratings. *Mathematical Finance*, 11(2):245–265.

Kingman, J. F. C. (1962). The imbedding problem for finite Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 1:14–24.

Norris, J. R. (1998). *Markov chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Seneta, E. (2016). Markov chains as models in statistical mechanics. *Statistical Science*, 31(3):399–414.

Tamaloussi, N. and Bouzaouit, A. (2020). Study of reliability in a repairable system by markov chains. *Acta Universitatis Sapientiae, Electrical and Mechanical Engineering*, 12(1):66–76.

# Estimation champ moyen pour un système excitateur/inhibiteur

Julien Chevallier [1] & Eva Löcherbach & Guilherme Ost

[1] *Université Grenoble Alpes, France, julien.chevallier1@univ-grenoble-alpes.fr*
[2] *Université Paris 1 Panthéon-Sorbonne, France, eva.locherbach@univ-paris1.fr*
[3] *Université fédérale de Rio de Janeiro, Brésil, guilhermeost@im.ufrj.br*

**Résumé.** Nous proposons une chaîne de Markov discrète pour modéliser les temps de décharge (*spikes*) de neurones. Le modèle comporte deux populations (excitatrice et inhibitrice) en interaction de type champ moyen dont le graphe d'interaction est un graphe d'Erdős-Rényi de paramètre $p \in [0,1]$. Le principal objectif est d'estimer cette probabilité de connexion $p$ en utilisant uniquement l'observation des *spikes*. La consistance de notre estimateur est prouvée dans la limite où le nombre de neurones et le temps d'observation tendent vers l'infini.

**Mots-clés.** Graphe d'interaction, Chaîne de Markov, Limite champ moyen

**Abstract.** A discrete Markov chain is proposed to model the spiking activity of neurons. Our model is structured with two populations (excitatory vs inhibitory) which are coupled via a mean field interaction and an Erdős-Rényi graph with parameter $p \in [0,1]$ as interaction graph. The main goal is to infer this connexion probability via the merely observation of the spiking times. Our estimator is proven to be consistant in the limit where both the number of neurons and the observation time goes to infinity.

**Keywords.** Interaction graph, Markov chain, Mean field limit

## 1 Introduction

Soit $N$ un entier positif qui représente le nombre de neurones dans le réseau. Nous allons considérer un modèle où chaque neurone est représenté par un processus à valeurs dans $\{0,1\}$: la valeur 1 encode une décharge et la valeur 0 encode l'absence de décharge. Nous supposons que l'ensemble des neurones $\{1,\dots,N\}$ est partitionné en deux sous-populations $\mathcal{P}_+$ et $\mathcal{P}_-$ (inconnues) qui ont, respectivement, un rôle excitateur et inhibiteur sur le système.

Notons $\theta = (\theta_{ij})_{i,j=1,\dots,N}$ la matrice d'adjacence d'un graphe d'Erdős-Rényi de taille $N$ et de paramètre inconnu $p \in [0,1]$. Plus précisément, les $N^2$ variables aléatoires $\theta_{ij}$ sont i.i.d. de loi $\mathrm{Ber}(p)$. Conditionnellement à $\theta$, nous considérons une chaîne de Markov $(X_t)_{t\in\mathbb{N}}$ à valeurs dans $\{0,1\}^N$ dont les probabilités de transition sont données par, pour tout $t \in \mathbb{N}^*$ et $x, y \in \{0,1\}^N$,

$$\mathbb{P}_\theta(X_t = y | X_{t-1} = x) = \prod_{i=1}^{N} (p_{\theta,i}(x))^{y_i} (1 - p_{\theta,i}(x))^{(1-y_i)}, \tag{1}$$

où $p_{\theta,i}(x)$ représente la probabilité que le $i$-ème neurone décharge au temps $t$ sachant que le système était dans la configuration $x$ au temps $t-1$. Nous considérons cette probabilité sous la forme suivante:

$$p_{\theta,i}(x) = \mu + \lambda \left( \frac{1}{N} \sum_{j \in \mathcal{P}_+} \theta_{ij} x_j + \frac{1}{N} \sum_{j \in \mathcal{P}_-} \theta_{ij} (1 - x_j) \right), \tag{2}$$

où $0 < \lambda < 1$ et $0 \leq \mu \leq 1 - \lambda$ sont deux paramètres inconnus.

La forme de l'Equation (1) indique que, conditionnellement à $X_{t-1} = x$, les coordonnées de $X_t$ sont indépendantes et de loi $\mathrm{Ber}(p_{\theta,i}(x))$. La forme de l'Equation (2) impose des interactions de type champ moyen avec deux populations: le neurone $i$ ressent la moyenne empirique des neurones $j \in \mathcal{P}_+$ d'un côté et $j \in \mathcal{P}_-$ de l'autre. De plus, comme les $\theta_{ij}$ sont positifs, l'influence des neurones $j \in \mathcal{P}_+$ est excitatrice: toutes choses égales par ailleurs, $p_{\theta,i}(x)$ est supérieure dans le cas où le $j$-ème neurone a déchargé ($x_j = 1$) par rapport au cas sans décharge ($x_j = 0$). Inversement, l'influence des neurones $j \in \mathcal{P}_-$ est inhibitrice.

Enfin, nous considérons que les proportions de neurones excitateurs $r_+^N = |\mathcal{P}_+|/N$ et inhibiteurs $r_-^N = |\mathcal{P}_-|/N$ convergent respectivement vers deux quantités **connues** $r_+$ et $r_-$ lorsque $N \to \infty$.

Le modèle considéré ici peut être vu comme un analogue à temps discret du modèle de Hawkes linéaire champ moyen étudié dans le papier de Delattre-Fournier (2016). La principale nouveauté de notre approche réside dans le fait que notre modèle considéré une population excitatrice **et** une population inhibitrice. En effet, le modèle de Hawkes linéaire ne permet pas de modéliser une population inhibitrice. De ce fait, une seule population est considérée dans le papier sus-cité.

# 2 Résultats

Tout d'abord, le fait que $\lambda < 1$ implique l'existence et l'unicité d'une version stationnaire de la chaîne décrite ci-dessus. Ceci est une conséquence assez immédiate de la représentation régénérative rétrograde que possède le modèle.

Soit $T$ un entier positif qui représente le temps d'observation. Supposons que nous observons une trajectoire $X_1, \ldots, X_T$ de la chaîne dans sa version stationnaire. Rappelons que le modèle comporte trois paramètres inconnus: $p$, $\mu$ et $\lambda$. Présentons maintenant les trois estimateurs que nous allons étudier.

Pour tout $t \in \mathbb{N}^*$, notons $\overline{X}_t = N^{-1} \sum_{i=1}^N X_t(i)$ la moyenne "spatiale" des observations, $Z_t = \sum_{s=1}^t X_s \in \mathbb{R}^N$ le vecteur des nombres de décharges entre $s = 1$ et $s = t$, et $\overline{Z}_t = N^{-1} \sum_{i=1}^N Z_t(i)$. Le premier estimateur n'est rien d'autre que la moyenne "spatio-temporelle" des observations,

$$\hat{m}_T = \frac{\overline{Z}_T}{T}.$$

Le second se base sur la variance "spatiale" des observations,

$$\hat{v}_T = \frac{(T+1)N}{T^3}\left[\frac{1}{N}\sum_{i=1}^{N}\left(Z_T(i)\right)^2 - \frac{T}{(T+1)}\left(\overline{Z}_T + \left(\overline{Z}_T\right)^2\right)\right].$$

Le dernier s'intéresse à la variance "temporelle" des observations,

$$\hat{w}_T^N = \frac{N}{T}\sum_{s=1}^{\lfloor T/\Delta \rfloor}\left(\overline{Z}_{s\Delta} - \overline{Z}_{(s-1)\Delta} - \Delta\hat{m}_T^N\right)^2,$$

où $\Delta$ est un entier à calibrer.

Une étude fine de la matrice aléatoire $\theta$ et de la dynamique du processus permet d'exprimer la limite des trois estimateurs en fonction des paramètres. Plus précisément, il existe une fonction explicite $\Psi : \mathbb{R}^3 \to \mathbb{R}^3$ (qui dépend de $r_+$) et une constante $K$ qui ne dépend que de $\mu, \lambda, p, r_+$ telles que

$$\mathbb{P}\left(\|(\hat{m}_T, \hat{v}_T, \hat{w}_T) - \Psi(\mu, \lambda, p)\| \geq \varepsilon\right) \leq \frac{K}{\varepsilon}\left(\frac{1}{\sqrt{N}} + \frac{\sqrt{N}}{\sqrt{T}}\right),$$

où l'on a choisit $\Delta$ de l'ordre de $\sqrt{T}$ pour minimiser la borne de droite. Enfin, comme la fonction $\Psi$ est inversible, le triplet $\Psi^{-1}(\hat{m}_T, \hat{v}_T, \hat{w}_T)$ donne un estimateur (faiblement) consistant de nos trois paramètres $(\mu, \lambda, p)$.

# Bibliographie

Delattre, S. et Fournier, N. (2016), Statistical inference versus mean field limit for Hawkes processes, *Electronic Journal of Statistics*, 10(1), pp. 1223-1295.

# La complexité d'échantillonnage des processus de décision markovien robuste est inférieure à celle des processus de décision markovien classique.

Pierre Clavier [1,2] & Laixi Shie [3] & Erwan Le Pennec [1]

[1] *CMAP, Ecole Polytechnique, pierre.clavier@poltechnique.edu ; erwan.le-pennec@polytechnique.edu ;*
[2] *INRIA Paris, HeKA*
[3] *California Institute of Technology, laixis@caltech.edu*

**Résumé.** Ce travail étudie la complexité de l'échantillonnage des processus de décision de Markov robustes (RMDP) L'objectif est d'optimiser les performances dans le pire des cas lorsque l'environnement se situe dans un ensemble d'incertitudes défini entourant un certain processus de Markov décisionnels (MDP) dit nominal. Malgré des efforts récents, la complexité d'échantillonnage des processus décisionnels de Markov robustes reste indéterminée. Bien que cette question ait été étudiée dans certains cas spécifiques, la généralisation des résultats existants reste incertaine, en particulier en comparaison avec les MDPs standards. En supposant l'accès à un modèle génératif qui échantillonne à partir du MDP nominal, nous examinons la complexité d'échantillonnage des RMDPs en utilisant une norme arbitraire comme fonction de "distance" pour l'ensemble d'incertitude, sous deux conditions couramment adoptées *sa*-rectangulaire et *s*-rectangulaire. Nous fournissons une borne supérieure quasi-optimale et une borne inférieure minimax correspondante pour les scénarios *sa*- rectangulaires. Pour les scénarios *s*-rectangulaires, nous améliorons la borne supérieure de pointe et dérivons une borne inférieure pour la norme $L_\infty$ et $L_1$ . Les résultats impliquent que les RMDPs peuvent être plus efficaces en termes d'échantillonnage que les MDP standards sous des normes générales dans les cas *sa*- et *s*-rectangulaires.

**Mots-clés.** Processus de Markov Décisionnels Robustes, complexité d'échantillonnage,

L'apprentissage par renforcement (RL) (Sutton, 1988) est un paradigme clé de l'apprentissage automatique, particulièrement remarquable pour son succès dans les applications pratiques. Le cadre de l'apprentissage par renforcement est souvent modélisé dans le contexte d'un processus de décision de Markov (MDP) et se concentre sur l'apprentissage de stratégies de prise de décision efficaces fondées sur des interactions avec un environnement. Cependant, les travaux de Mannor et al. (2004) ont mis en évidence une vulnérabilité du RL, révélant sa sensibilité aux erreurs d'estimation dans les probabilités de récompense et de transition. Un exemple typique des problèmes rencontrés par le RL est lorsque qu'en raison d'un écart entre les simulations et les applications réelles (dit sim-to-real), les politiques apprises dans des environnements idéalisés et simulés peuvent échouer de manière catastrophique lorsqu'elles sont déployées dans des environnements avec de légers changements ou des perturbations adverses (Klopp et al., 2017; Mahmood et al., 2018).

Pour résoudre ce problème, les MDP robustes (RMDP), proposés par Iyengar (2005) et Nilim and El Ghaoui (2005), ont fait l'objet d'une attention considérable. Les RMDP sont

formulés comme des problèmes max-min, recherchant des politiques qui résistent aux erreurs d'estimation du modèle dans un ensemble d'incertitudes spécifié. Malgré les avantages de la robustesse, la résolution des RMDP est NP-hard pour les ensembles d'incertitude généraux (Nilim and El Ghaoui, 2005). Pour surmonter cette difficulté, l'hypothèse de rectangularité des ensemble d'incertitude est souvent adoptée, les ensembles d'incertitude sont ainsi structurés comme des produits de sous-ensembles indépendants pour chaque état ou paire état-action, désignés par les hypothèses $s$-rectangulaire ou $sa$-rectangulaire (voir les définitions (5) et (7)). Ces deux hypothèses facilitent l'utilisation de méthodes telles que l'itération robuste de la valeur et l'itération robuste de la politique, en préservant de nombreuses propriétés structurelles des MDP (Ho et al., 2021). Les ensembles $s$-rectangulaires, bien que moins restrictifs, posent de plus grands défis, tandis que les ensembles $sa$-rectangulaires permettent des politiques déterministes apparentées aux MDP non robustes (Wiesemann et al., 2013). Enfin, il est important de noter que, si l'incertitude de la récompense peut facilement être gérée, il est plus difficile de gérer l'incertitude du noyau de transition, (Kumar et al., 2022; Derman et al., 2021).

La question de l'efficacité de l'échantillonnage est centrale dans les problèmes de RL allant de la pratique à la théorie. Bien que les bornes minimax soient atteintes dans les travaux de Azar et al. (2013); Li et al. (2023) dans le contexte des MDP classiques, cet objectif n'est pas encore atteint en général, dans le contexte des RMDP. Plus précisément, il existe des travaux antérieurs sur la complexité de l'échantillon de RL robuste pour quelques divergences spécifiques telles que la variation totale $TV$, $L_p$, $\chi^2$, $KL$, et Wasserstein (Yang et al., 2022a; Zhou et al., 2021; Panaganti and Kalathil, 2022), alors que de tels résultats restent incertains pour des classes plus générales de divergences. À ce jour, à notre connaissance, les résultats de la complexité de l'échantillon qui atteignent l'optimalité minimax pour tout rayon d'incertitude sont limités à un seul cas à savoir la distance $TV$ dans le cas $sa$-rectangulaire. (Shi et al., 2023).

Dans ce travail, nous nous concentrons sur la question de la complexité d'échantillonnage des RMDP avec une norme arbitraire. Cette généralisation est intéressante à la fois en pratique et en théorie. En pratique, de nombreuses applications reposent des approches qui impliquent des normes arbitraires et ad-hoc, différentes de celles qui ont déjà été étudiées théoriquement. Par exemple, le contrôle robuste peut être spécifique à une tâche, utilisant une distance de Mahalanobis (Jiang and Zhang, 2018) pour construire les ensembles d'incertitude. D'un point de vue théorique, il est intéressant d'étudier le coût statistique de la robustesse en RL dans des scénarios plus généraux, ce qui conduit à deux questions ouvertes auxquelles nous essayerons de répondre. L'une d'entre elles concerne sur la complexité d'échantillonnage pour la résolution du RL robuste par rapport à la résolution de RL classique. En particulier, pour le cas spécifique de la distance $TV$, Shi et al. (2023) a montré que la complexité d'échantillonnage pour résoudre le RL robuste est au plus la même et parfois (lorsque le niveau d'incertitude est relativement grand) plus petite que celle de du RL standard. Cela motive la question ouverte suivante :

*Le RL robuste est-il plus efficace en termes d'échantillons que le RL standard pour des normes générales ?*

Une seconde question concerne les comparaisons entre la complexité d'échantillonnage de

la résolution des RMDPs $s$-rectangulaires et celle de la résolution de RMDPs avec l'hypothèse de $sa$-rectangularité. On peut souligner que les RMDPs $s$-rectangulaires ont des formulations d'optimisation plus compliquées avec des variables supplémentaires (niveaux d'incertitude pour chaque action) à optimiser. Cela conduit à une classe plus riche de politique optimale, à savoir des politiques stochastiques dans les cas $s$-rectangulaires, contrairement à la classe des politiques déterministes pour les cas $sa$-rectangulaires. En outre, la limite supérieure de la complexité d'échantillonnage existante pour la résolution des RMDP $s$-rectangulaires est plus grande que celle de la résolution des RMDP $sa$-rectangulaires (Yang et al., 2022a) pour les cas étudiés. Cela motive la questions suivante:

*La résolution de s-rectangulaires RMDPs nécessite-t-elle, en effet, plus d'échantillons que la résolution de sa-rectangulaires RMDPs avec des normes générales ?*

**Contributions.** Dans ce travail, nous abordons chacune des deux questions discutées ci-dessus. En particulier, nous fournissons la première analyse de complexité d'échantillon pour les RMDP avec des normes générales sous les conditions de $s$- et $sa$-rectangularité. Par commodité, nous présentons une comparaison détaillée de l'état de l'art existant et nos résultats dans le tableau 1 et discutons des contributions et de leurs implications ci-dessous. En ce qui concerne la première question, nous illustrons nos résultats dans les cas $sa$- et $s$-rectangulaire dans la Figure 1. Dans le cas de la $sa$-rectangularité, nous dérivons une borne supérieure de complexité d'échantillon pour les RMDP en utilisant des normes générales (Théorème 2.1) de l'ordre de :

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right). \tag{1}$$

De plus, nous fournissons une borne inférieure minimax correspondante (Théorème 2.2) qui confirme la quasi-optimalité de la borne supérieure pour la quasi-totalité de la plage du niveau d'incertitude. Cela correspond à la complexité d'échantillonnage quasi-optimale dérivée dans Shi et al. (2023) pour le cas spécifique de la distance $TV$ et $sa$ rectangulaire, tout en étant valable pour n'importe quelle norme arbitraire. Dans le cas d'une $s$-rectangularité, nous fournissons une borne supérieure de complexité pour la résolution des RMDP avec n'importe quelle norme générale de l'ordre de :

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right).$$

Ce résultat améliore l'état de l'art antérieur $\widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right)$ dans Clavier et al. (2023) pour le cas spécifique de $L_p$ lorsque $\tilde{\sigma} \lesssim 1-\gamma$ par au moins un facteur de $1/(1-\gamma)$. De plus, nous présentons une borne inférieure pour un cas représentatif avec la norme $L_\infty$, qui atteint la borne supérieure. À notre connaissance, il s'agit de la première borne inférieure pour la résolution de RMDPs avec $s$-rectangularité. Enfin, nous sommes en mesure de jeter un nouvel éclairage sur la seconde question grâce à nos nouveaux résultats. En particulier, comme l'illustre la figure 1, nos résultats mettent en évidence le fait que le RL robuste est au moins aussi efficace que le RL standard pour les normes générales, et qu'il peut parfois l'être davantage. Ce résultat a une importance pratique considérable et constitue une motivation

| Résultat | Reference | Distance | *sa*-rectangulaire | | *s*-rectangulaire | |
|---|---|---|---|---|---|---|
| | | | $0 < \sigma \lesssim 1-\gamma$ | $1-\gamma \lesssim \sigma < \sigma_{\max}$ | $0 < \sigma \lesssim 1-\gamma$ | $1-\gamma \lesssim \sigma < \sigma_{\max}$ |
| Borne supérieure | Yang et al. (2022b) | TV | $\frac{S^2A(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$ | $\frac{S^2A(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$ | $\frac{S^2A^2(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$ | $\frac{S^2A^2(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$ |
| | Panaganti and Kalathil (2022) | TV | $\frac{S^2A}{(1-\gamma)^4\varepsilon^2}$ | $\frac{S^2A}{(1-\gamma)^4\varepsilon^2}$ | × | × |
| | Shi et al. (2023) | TV | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$ | × | × |
| | Clavier et al. (2023) | $L_p$ | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{(1-\gamma)^4\varepsilon^2}$ | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{(1-\gamma)^4\varepsilon^2}$ |
| | **Ce travail** | $\|\cdot\|$ | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$ | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$ |
| Borne inférieure | Yang et al. (2022b) | TV | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA(1-\gamma)}{\sigma^4\varepsilon^2}$ | × | × |
| | Shi et al. (2023) | TV | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$ | × | × |
| | **Ce travail** | $L_\infty$ | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$ | $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ | $\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$ |

Table 1: Comparaisons avec des résultats antérieurs (aux termes logarithmiques près) de la complexité nécessaire pour obtenir la recherche une politique optimale à $\varepsilon$ près pour un processus de décision de Markov robuste, où $\sigma$ est le rayon de l'ensemble d'incertitude et $\sigma_{\max}$ défini dans 2.1.

essentielle pour l'utilisation et l'étude de la robustesse. Plus précisément, le RL robuste ne réduit pas seulement la vulnérabilité du RL aux erreurs d'estimation et aux écarts entre les simulations et le réel, mais conduit également à une meilleure efficacité en termes de complexité d'échantillonnage. En termes de comparaison des implications statistiques de la *sa*- et de la *s*- rectangularité, nos résultats montrent que la résolution de RMDPs *s*-rectangulaires n'est pas plus difficile que la résolution de RMDPs *sa*-rectangulaires en termes d'exigence d'échantillon (voir le théorème 2.3)

# 1 Formulation du problème : Processus de décision de Markov robustes

**Processus de décision de Markov (MDP) standard.** Un MDP actualisé à horizon infini est représenté par $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, r)$, où $\mathcal{S} = \{1, \cdots, S\}$ et $\mathcal{A} = \{1, \cdots, A\}$ sont les espaces d'état et d'action finis, respectivement, $\gamma \in [0, 1)$ est le facteur d'actualisation, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ représente le noyau de transition des probabilités, et $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ est la fonction de récompense immédiate, qui est supposée être déterministe. De plus, nous supposons que la fonction de récompense est bornée en $(0, 1)$ sans perte de généralité. La politique que nous recherchons est définie par $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, qui spécifie la probabilité de sélection d'action sur l'espace d'action pour tous les états. Enfin, pour caractériser la récompense cumulative, la fonction de valeur $V^{\pi,P}$ pour toute politique $\pi$ sous le noyau de transition $P$ est définie par $\forall s \in \mathcal{S}$

$$V^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s\right]. \tag{2}$$

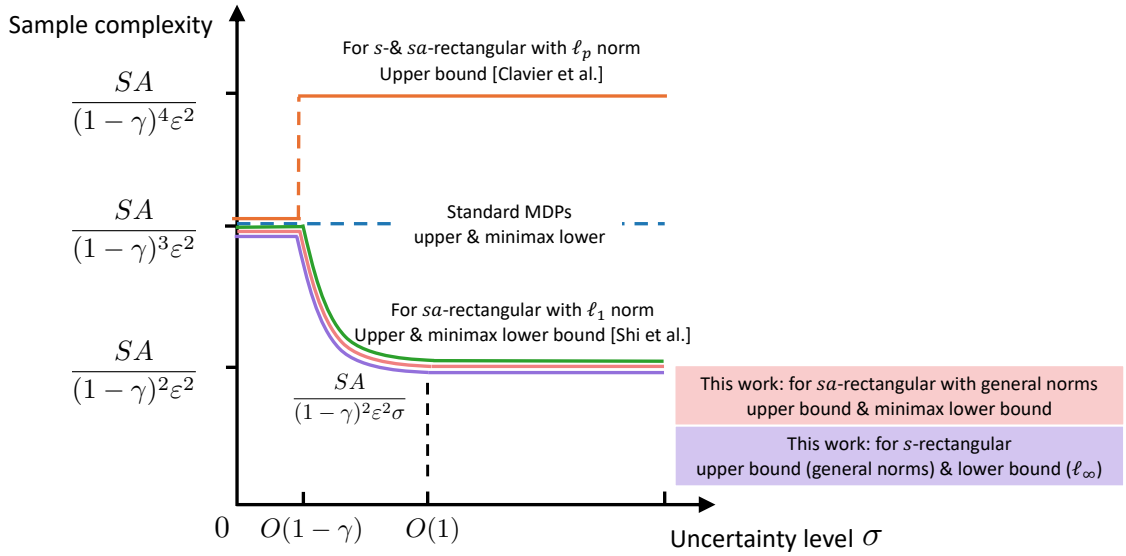



Figure 1: Les résultats de la complexité de l'échantillon pour les RMDP avec *sa*- et *s*-rectangularité avec des normes générales et des comparaisons avec les arts antérieurs (Shi et al., 2023) (pour la norme $\ell_1$, ou appelée distance de variation totale) et (Clavier et al., 2023) (pour la norme $L_p$ avec $1 \leq p \leq \infty$).

L'espérance est prise sur le caractère aléatoire de la trajectoire $\{s_t, a_t\}_{t=0}^{\infty}$ générée par l'exécution de la politique $\pi$ sous le noyau de transition $P$, de sorte que $a_t \sim \pi(\cdot \,|\, s_t)$ et $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ pour tout $t \geq 0$. De la même manière, la Q-fonction $Q^{\pi,P}$ associée à toute politique $\pi$ sous le noyau de transition $P$ comme : $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ : est définie comme suit

$$Q^{\pi,P}(s,a) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\middle|\, s_0, a_0 = s, a\right], \tag{3}$$

avec une espérance prise sur le caractère aléatoire de la trajectoire sous la politique $\pi$.

**RMDPs robustes du point de vue de la distribution** Nous considérons des MDP robustes sur le plan de la distribution (RMDP) dans le cadre d'un horizon infini actualisé, dénotés par $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^{\sigma}_{\|.\|}(P^0), r\}$, où $\mathcal{S}, \mathcal{A}, \gamma, r$ sont les mêmes ensembles et paramètres que dans les PDM standard. La principale différence par rapport aux MDP standard est qu'au lieu d'utiliser un noyau de transition fixe $P$, il permet au noyau de transition d'être choisi arbitrairement dans un ensemble d'incertitude prescrit $\mathcal{U}^{\sigma}_{\|.\|}(P^0)$ centré autour d'un noyau *nominal* $P^0 : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, où l'ensemble d'incertitude est spécifié en utilisant une métrique de distance $\|.\|$ de rayon $\sigma > 0$. Cette définition est générale et inclut $TV$, $L_p, p > 1$ et toute norme, telle que la distance de Mahalanobis, par exemple. Cependant, elle n'inclut pas les divergences telles que $KL$ et $\chi^2$. En particulier, étant donné le noyau de transition nominal $P^0$ et un certain niveau d'incertitude $\sigma$, l'ensemble d'incertitude–avec une norme arbitraire $\| \, \| : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^+$ ou de $\mathbb{R}^{\mathcal{SA}}$ dans le cas *s*-rectangulaire, est spécifié par

$\mathcal{U}_{\|.\|^\sigma(P^0):=\times_{s,a}\,\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a})}$

$$\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \left\|P_{s,a} - P^0_{s,a}\right\| \leq \sigma\right\}, \tag{4}$$

où nous désignons un vecteur du noyau de transition $P$ ou $P^0$ au couple état-action $(s,a)$ respectivement par

$$P_{s,a} \coloneqq P(\cdot \mid s,a) \in \mathbb{R}^{1\times S}, \quad P^0_{s,a} \coloneqq P^0(\cdot \mid s,a) \in \mathbb{R}^{1\times S}. \tag{5}$$

En d'autres termes, l'incertitude est imposée de manière découplée pour chaque paire état-action, obéissant à la soi-disant *sa*-rectangularité (Zhou et al., 2021; Wiesemann et al., 2013). Dans ce travail, nous considérerons toute norme arbitraire définie comme $\|.\|$. Plus généralement, nous définissons les RMPD $s$-rectangulaires comme $\mathcal{U}^{\sigma}_{\|.\|}(P) = \otimes_s \mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P_s)$, pour la norme arbitraire $\|.\|$, à l'aide de la définition suivante

$$P_s \coloneqq P(\cdot,\cdot \mid s) \in \mathbb{R}^{1\times SA}, \quad P^0_s \coloneqq P^0(\cdot,\cdot \mid s) \in \mathbb{R}^{1\times SA}. \tag{6}$$

L'incertitude est imposée de manière découplée pour chaque paire d'états, et un budget fixe donné à un état pour toutes les actions est défini. Pour obtenir une signification similaire pour le rayon de la boule entre les hypothèses *sa*-rectangulaire et *s*-rectangulaire, nous devons renormaliser le rayon en fonction de la norme comme dans Yang et al. (2022a). L'ensemble d'incertitude $s$ est alors défini en utilisant le rayon renormalisé $\tilde{\sigma}$ comme suit

$$\mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P_s) \coloneqq \left\{P'_s \in \Delta(\mathcal{S})^{\mathcal{A}} : \|P'_s - P_s\| \leq \tilde{\sigma} = \sigma \|1\| \right\}, \tag{7}$$

où 1 représente le vecteur unitaire. Dans le cas spécifique des normes $L_1$, $L_p$ et $L_\infty$, $\tilde{\sigma}$ est égal à $|\mathcal{A}|, |\mathcal{A}|^{1/p}$ et 1. Notez que cette échelle permet une comparaison équitable entre les PDM *sa*- et *s*-rectangulaires. Dans les RMDP, nous nous intéressons à la performance la plus défavorable d'une politique $\pi$ sur tous les noyaux de transition possibles dans l'ensemble d'incertitude. Cette performance est mesurée par la fonction de valeur robuste $V^{\pi,\sigma}$ et la fonction Q robuste $Q^{\pi,\sigma}$ dans $\mathcal{M}_{\mathsf{rob}}$, définies respectivement comme $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$

$$V^{\pi,\sigma}(s) \coloneqq \inf_{P\in\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0)} V^{\pi,P}(s), \tag{8}$$

$$Q^{\pi,\sigma}(s,a) \coloneqq \inf_{P\in\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0)} Q^{\pi,P}(s,a). \tag{9}$$

De la même manière, nous définissons la fonction de valeur de $s$-rectangularité.

$$V^{\pi,\sigma}_s(s) \coloneqq \inf_{P\in\mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P^0)} V^{\pi,P}(s), \tag{10}$$

**Noyau nominal empirique.** Le noyau empirique de transition nominale $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$ peut être construit sur la base de la fréquence empirique des transitions d'état, c'est-à-dire ,$\forall (s,a) \in \mathcal{S}\times\mathcal{A}$

$$\widehat{P}^0(s' \mid s,a) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left\{s_{i,s,a} = s'\right\}, \tag{11}$$

qui mène au RMDP empirique $\widehat{\mathcal{M}}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^{\sigma}_{\|.\|}(\widehat{P}^0), r\}$. Toutes les quantités définies avec opérateur $\hat{.}$ sont définies comme précédemment, mais dans le MRPD empirique comme les Q fonction ou fonction de valeur empirique $\hat{Q}$ et $\hat{V}$.

# 2 Garanties théoriques

## 2.1 Ensemble d'incertitude *sa*-rectangulaire avec normes générales

Pour commencer, nous considérons les RMDPs avec *sa*-rectangularité avec des normes arbitraires. Nous commençons par fournir la limite supérieure de complexité d'échantillon.

**Theorem 2.1 (Borne supérieure pour l'hypothèse *sa*-rectangulaire.)** *Nous considérons l'ensemble d'incertitude $\mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\cdot)$ associé à la norme $\|\cdot\|$ et notons $\sigma_{\max} \coloneqq \max_{p_1,q_1\in\Delta(\mathcal{S})} \|p_1 - p_2\|$ le rayon maximal. Pour un niveau de confiance $\delta \in (0,1)$, un facteur d'actualisation $\gamma \in \left[\frac{1}{4}, 1\right)$, et un rayon $\sigma \in (0, \sigma_{\max}]$ nous définissons la politique oracle dans le MDP empirique $\widehat{\pi}$ qui est le résultat du problème d'optimisation dans le RMPDS empirique avec une erreur $\varepsilon_{\mathsf{opt}}$ tel que $\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \leq \varepsilon_{\mathsf{opt}}$. Avec probabilité au moins $1-\delta$, on a*

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq \varepsilon + \frac{7\varepsilon_{\mathsf{opt}}}{1-\gamma} \tag{12}$$

*pour tout $\varepsilon \in (0, \sqrt{1/\max\{1-\gamma,\sigma\}}]$, si le nombre d'échantillons obéit*

$$NSA \gtrsim \frac{C_1 SA}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}$$

*avec $C_1$ une constante universelle positive.*

Nous introduisons la borne inférieure minimax-optimale suivante pour les normes générales afin de vérifier de l'intérêt de la borne supérieure ci-dessus.

**Theorem 2.2 (Borne inférieure pour l'hypothèse *sa*-rectangulaire)** *En considérant l'ensemble d'incertitude $\mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\cdot)$ associé à une norme arbitraire $\|\cdot\|$ et le rayon maximal $\sigma_{\max} \coloneqq \max_{p_1,q_1\in\Delta(\mathcal{S})} \|p_1 - p_2\|$ nous définissons le tuple $(S, A, \gamma, \sigma, \varepsilon)$, avec $\gamma \in \left[\frac{1}{2}, 1\right)$, $\sigma \in (0, \sigma_{\max}(1-c_0)]$ et $0 < c_0 \leq \frac{1}{8}$ une constante positive et $\varepsilon \in \left(0, \frac{c_0}{256(1-\gamma)}\right]$. Nous pouvons contruire deux RMPDs $\mathcal{M}_0, \mathcal{M}_1$ tel que ayant donné un jeux de données avec $N$ échantillons indépendants pour chaque couple état-action échantillonné du MDP nominal (pour chaque $\mathcal{M}_0$ ou $\mathcal{M}_1$ respectivement), on a*

$$\inf_{\widehat{\pi}} \max_{\mathcal{M}\in\{\mathcal{M}_0,\mathcal{M}_1\}} \left\{ \mathbb{P}_{\mathcal{M}}\Big( \max_{s\in\mathcal{S}} \left[V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s)\right] > \varepsilon \Big) \right\} \geq \frac{1}{8},$$

*tant que*

$$NSA \leq \frac{C_2 SA}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}.$$

*Ici $C_2$ est une constante universelle positive, l'infimum est pris sous tous les estimateurs $\widehat{\pi}$, et $\mathbb{P}_0$ (respectivement. $\mathbb{P}_1$) dénote la probabilité lorsque le RMDP est $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).*

## 2.2 Ensemble d'incertitude $s$-rectangulaire avec normes générales

Pour continuer, nous passons au cas où l'ensemble d'incertitude est construit sous $s$-rectangularité. Le théorème suivant présente la limite supérieure de la complexité de l'échantillon pour l'apprentissage d'une politique optimale *varepsilon* pour les RMDP avec $s$-rectangularité.

**Theorem 2.3 (Borne supérieure pour l'hypothèse $s$-rectangulaire.)** *Nous considérons ici l'ensemble d'incertitude $\mathcal{U}^{\mathbf{s},\widetilde{\sigma}}_{\|\cdot\|}(\cdot)$ sous l'hypothèse s-rectangulaire. Nous définissons également le facteur d'actualisation $\gamma \in \left[\frac{1}{4}, 1\right)$, le rayon d'incertitude $\tilde{\sigma} = \sigma \|1\|$ et le niveau de confiance $\delta \in (0,1)$. Nous définissons la politique oracle dans le MDPs empirique $\widehat{\pi}$ qui est le résultat du problème d'optimisation dans le RMPDS empirique avec une erreur $\varepsilon_{\mathsf{opt}}$ tel que $\widehat{V}^{\widehat{\pi}\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \leq \varepsilon_{\mathsf{opt}}$. Avec probabilité au moins $1-\delta$, on a*

$$\forall s \in \mathcal{S} : \quad V^{\star,\widetilde{\sigma}}(s) - V^{\widehat{\pi},\widetilde{\sigma}}(s) \leq \varepsilon + \frac{7\varepsilon_{\mathsf{opt}}}{1-\gamma}$$

*tant que le nombre d'échantillons obéit*

$$NSA \gtrsim \frac{C_3 SA}{(1-\gamma)^2\varepsilon^2} \min\left\{ \frac{1}{\max\{1-\gamma,\sigma\}}, \frac{1}{\sigma \min_{s\in\mathcal{S}} \left\{ \|\pi^*_s\|_* \|1\|, \|\hat{\pi}_s\|_* \|1\| \right\}} \right\}, \tag{13}$$

*avec $C_3$ une constante positive universelle.*

où $\hat{\pi}_s \in \Delta_A$ dénote la politique des RMPD empiriques à l'état $s$, $\pi^*_s \in \Delta_A$ la politique optimale à l'état $s$ et $\|.\|_*$ la norme duale. De plus, nous fournissons les bornes inférieures pour les normes $L_\infty$ et $L_1$ dans le théorème suivant :

**Theorem 2.4 (Borne inférieure pour l'hypothèse $s$-rectangulaire.)** *Ici, le théorème est valable pour la norme infinie $L_\infty$ et la norme $L_1$. En considérant l'ensemble d'incertitude $\mathcal{U}^{\mathbf{s},\widetilde{\sigma}}_{\|\cdot\|}(\cdot)$ associé à une norme arbitraire $\|\cdot\|$ et le rayon maximal $\sigma_{\max} \coloneqq \max_{p_1,q_1\in\Delta(\mathcal{S})} \|p_1 - p_2\|$ nous définissons le tuple $(S, A, \gamma, \sigma, \varepsilon)$, avec $\gamma \in \left[\frac{1}{2}, 1\right)$, $\sigma \in (0, \sigma_{\max}(1-c_0)]$ et $0 < c_0 \leq \frac{1}{8}$ une constante positive et $\varepsilon \in \left(0, \frac{c_0}{256(1-\gamma)}\right]$. Ici*

*Nous pouvons construire deux RMPDs $\mathcal{M}_0, \mathcal{M}_1$ tel que étant donné un jeu de données avec $N$ échantillons indépendants pour chaque couple état-action échantillonné du MDP nominal (pour chaque $\mathcal{M}_0$ ou $\mathcal{M}_1$ respectivement), on a*

$$\inf_{\widehat{\pi}} \max_{\mathcal{M}\in\{\mathcal{M}_0,\mathcal{M}_1\}} \left\{ \mathbb{P}_{\mathcal{M}}\Big( \max_{s\in\mathcal{S}} \left[ V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \right] > \varepsilon \Big) \right\} \geq \frac{1}{8},$$

*tant que*

$$NSA \leq \frac{C_2 SA}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}.$$

*Ici $C_2$ est une constante universelle positive, l'infimum est prit sous tous les estimateurs $\widehat{\pi}$, et $\mathbb{P}_0$ (respectivement. $\mathbb{P}_1$) dénote la probabilité lorsque le RMDP est $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).*

# 3 Conclusion

Ce travail a fait progresser le domaine en affinant les limites de la complexité de l'échantillon pour apprendre des processus décisionnels de Markov robustes lorsque l'ensemble d'incertitude est caractérisé par une norme arbitaire en supposant la présence d'un modèle génératif. Nos résultats renforcent non seulement le corpus de connaissances existant en améliorant les limites supérieures et inférieures, mais soulignent également que l'apprentissage des MDP $s$-rectangulaires est moins difficile en termes de complexité d'échantillonnage que les MDP classiques $sa$-rectangulaires. Ce travail représente un effort considérable pour fournir des résultats avec une borne minimax, car les résultats précédents concernant les cas $s$-rectangulaires n'étaient pas minimax optimaux. En outre, nous avons établi la complexité d'échantillonnage minimax pour les RMDP en utilisant une norme arbitraire, en démontrant qu'elle n'est jamais plus grande que celle requise pour l'apprentissage des MDP standard. Notre recherche fournit des pistes potentielles pour des travaux futurs, comme l'exploration de la caractérisation de la complexité d'échantillonnage pour les RMDP dans une famille plus large d'ensembles d'incertitude, tels que ceux définis par la $f$-divergence. Il serait souhaitable de disposer d'une base théorique plus unifiée, puisque la distance entre les mesures de probabilité est plus naturelle à définir à l'aide de divergences. De plus, il serait intéressant de se concentrer également sur la question de l'horizon fini et le cadre linéaire. Une telle extension contribuerait à une compréhension plus complète des cas tabulaires.

# References


Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

Clavier, P., Pennec, E. L., and Geist, M. (2023). Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*.

Derman, E., Geist, M., and Mannor, S. (2021). Twice regularized MDPs and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34.

Ho, C. P., Petrik, M., and Wiesemann, W. (2021). Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280.

Jiang, C. and Zhang, S.-B. (2018). A novel adaptively-robust strategy based on the mahalanobis distance for gps/ins integrated navigation systems. *Sensors*, 18(3):695.

Klopp, O., Lounici, K., and Tsybakov, A. B. (2017). Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564.

Kumar, A., Levine, A., Goldstein, T., and Feizi, S. (2022). Certifying model accuracy under distribution shifts. *arXiv preprint arXiv:2201.12440*.

Li, G., Yan, Y., Chen, Y., and Fan, J. (2023). Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278*.

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., and Bergstra, J. (2018). Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. (2004). Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, page 72.

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798.

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.

Yang, W., Zhang, L., and Zhang, Z. (2022a). Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248.

Yang, W., Zhang, L., and Zhang, Z. (2022b). Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR.

# Copules

# Procédure de test d'hypothèses composites pour l'analyse jointe de séries de probabilités critiques.

Annaïg De Walsche [1,2,*] & Franck Gauthier [2] & Alain Charcosset [2] & Tristan Mary-Huard [1,2]

[1] *UMR MIA Paris-Saclay, INRAE, AgroParisTech, Université Paris-Saclay, 91120 Palaiseau, France*

[2] *UMR Génétique Quantitative et Evolution - Le Moulon, INRAE, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France*

[*] *Corresponding author: annaig.de-walsche@inrae.fr*

**Résumé.** L'analyse jointe de résultats de différentes expériences pour identifier des configurations complexes est un objectif typique de l'intégration de données. On considère ici le cas d'une collection d'éléments $i = 1, ..n$ (par exemple des gènes) pour lesquels les hypothèses $H_{0i}^q$ : « l'élément $i$ n'a pas d'effet dans la condition $q$ » ont été testées pour $Q$ conditions. Chaque observation $i$ consiste donc en un vecteur de $Q$ probabilités critiques. L'objectif de l'analyse est alors d'identifier les éléments qui ont un effet dans toutes les conditions ou dans un sous-ensemble prédéfini de conditions. Les probabilités critiques doivent alors être combinées de manière flexible afin d'explorer des hypothèses complexes (appelées hypothèses composites), tout en contrôlant le taux de faux positif. Nous proposons une procédure de test d'hypothèses composites utilisant un modèle de mélange multivarié où chaque $Q$-uplet de probabilités critiques appartient à une des $2^Q$ classes caractérisée par une combinaison spécifique d'états de $H_0^q$ et $H_1^q$. Notre méthode prend en compte la structure de dépendance entre les $Q$ probabilités critiques, qui est modélisée dans les lois jointes conditionnelles à l'aide d'une fonction copule. L'inférence de ce modèle de mélange à $2^Q$ composantes est réalisée efficacement permettant son application à des cas où le nombre de marqueurs est en $\mathcal{O}(10^5)$, et où $Q = 20$. Elle consiste en deux étapes indépendantes : tout d'abord l'ajustement d'un modèle de mélange non paramétrique sur la distribution marginale de chacune des $Q$ séries de probabilités critiques, puis l'estimation des proportions des composantes du modèle de mélange et des paramètres de copule via un algorithme EM. L'étape (E) est optimisée pour limiter l'empreinte mémoire de la procédure, passant de $O(n \times 2^Q)$ à $O(n + 2^Q)$. Des applications sur des données simulées ont été réalisées donnant des résultats concluants tant en termes de contrôle de faux positif et de puissance de détection qu'en terme d'efficacité de la méthode (temps de calcul et gestion de la mémoire). L'intérêt de la méthode est illustré par une analyse conjointe d'études d'association génétique afin de détecter des gènes pléiotropes parmi un ensemble de 14 troubles psychiatriques.

**Mots-clés.** Hypothèse composite, modèle de mélange, tests multiples, intégration des données, pléiotropie.

**Abstract.** Data integration often involves analysing results from different experiments to identify complex patterns. In this context, we consider a scenario where we have a collection of elements $i = 1, ...n$ (genes, for example) for which the hypotheses $H_{0i}^q$: "element $i$ has

no effect in condition $q$" have been tested for $Q$ conditions. Each observation $i$, therefore, consists of a vector of $Q$ critical probabilities. The analysis aims to identify the elements that have an effect in all the conditions or a predefined subset of those conditions. The critical probabilities must then be combined flexibly to explore complex hypotheses (called composite hypotheses) while controlling the false positive rate. To achieve this, we need to combine the critical probabilities in a flexible way to explore complex hypotheses (called composite hypotheses) while controlling the false positive rate. We propose a composite hypothesis testing procedure based on a model where the $Q$-uplet of p-value associated with each gene/marker is distributed as a multivariate mixture where each of the $2^Q$ components corresponds to a specific combination of $H_0^q$ and $H_1^q$ states. Our method explicitly accounts for the dependence structure across p-value series through a copula function. The inference of this $2^Q$ component mixture model is performed efficiently, allowing its application to cases where the number of markers is $\mathcal{O}(10^5)$, and where $Q = 20$. The inference procedure consists of two independent steps: first, fitting a non-parametric mixture model to the marginal distribution of each $Q$ series of p-values, then estimating the proportions of the mixture model components and the copula parameters using an EM algorithm. Step (E) is optimised to reduce the memory burden of the procedure from $O(n \times 2^Q)$ to $O(n + 2^Q)$. Applications on simulated data have been carried out, with conclusive results regarding false positive control and detection power and the method's efficiency (computation time and memory management). The interest in the method is illustrated by a joint analysis of genetic association studies to detect pleiotropic genes among 14 psychiatric disorders.

**Keywords.** Composite hypothesis, mixture model, multiple testing, data integration, pleiotropy.

# 1 Introduction

Considérons une étude dont l'objectif est d'évaluer l'effet conjoint d'un traitement sur deux tissus différents. On cherche alors à définir une procédure de test qui rejette l'hypothèse $H_0$ "le traitement n'a pas d'effet conjoint", lorsque les deux hypothèses $H_0^1$ "le traitement n'a pas d'effet sur le tissu 1" et $H_0^2$ "le traitement n'a pas d'effet sur le tissu 2" sont fausses. Cela correspond à un cas particulier de test d'hypothèse composite où l'hypothèse composite $H_0$ à tester est $H_0^1 \cup H_0^2$. Une approche courante pour effectuer des tests d'hypothèses composites (THC) consiste à combiner les statistiques de test et/ou les probabilités critiques dérivées pour chacune des hypothèses marginales $H_0^1$ et $H_0^2$ en une seule statistique globale. Si les probabilités critiques associées à $H_0^1$ et $H_0^2$ peuvent usuellement être obtenues à l'aide de procédures statistiques classiques, construire une statistique de test adéquate (avec une distribution connue sous $H_0$) ainsi qu'une procédure de test valide (garantissant le contrôle du taux de faux positifs au niveau nominal requis) pour le test de l'hypothèse composite $H_0$ n'est pas trivial. Le test d'hypothèse composite de la forme $H_0^1 \cup H_0^2$ a été étudié dès le début des années 80, avec les travaux de [13], et son extension au cas $H_0^1 \cup ... \cup H_0^Q$ avec $Q \geq 2$ a été explorée par [2]. En génétique, le THC peut être utilisé pour analyser conjointement les résultats issus de plusieurs analyses d'association, réalisées à partir de panels non disjoints

(i.e. une partie des individus est commune aux différents panels). ([5, 16, 8]). Dans un tel cas le THC est réalisé au niveau du marqueur, ce qui entraîne un grand nombre d'hypothèses composites testées simultanément. Par ailleurs les différentes probabilités critiques collectées pour un même marqueur ne peuvent être considérées comme indépendantes du fait de la présence d'individus communs à tous les panels.

Nous proposons une approche de THC basée sur un modèle où le $Q$-uplet de probabilités critiques associé à chaque marqueur est issue d'un mélange multivarié où chacune des $2^Q$ composantes correspond à une combinaison spécifique d'états $H_0^q$ et $H_1^q$. La méthode, appelée **qch_copula**, prend en compte la structure de dépendance entre les séries de probabilités critiques à l'aide d'une fonction copule. Nous montrons comment l'inférence d'un tel modèle de mélange à $2^Q$ composantes peut être réalisée efficacement, permettant son application à des cas où le nombre de marqueurs est en $\mathcal{O}(10^5)$, et où $Q = 20$. La procédure est illustrée sur des données simulées, ainsi que sur un exemple de détection de gènes pléiotropes parmi un ensemble de 14 troubles psychiatriques en analysant conjointement des études d'association génétiques. Les performances obtenues en termes de puissance de détection et de contrôle du taux d'erreur de type I sont très supérieures à celles des méthodes concurrentes.

# 2 Hypothèse composite

On considère une collection d'éléments $i = 1, ...n$ (par exemple des gènes ou des SNP) dont l'effet a été testé dans $q = 1, ..., Q$ conditions. Nous désignons par $H_0^q$ (resp. $H_1^q$ ) l'hypothèse nulle (resp. alternative) correspondant au test $q$ ($1 \leq q \leq Q$) et considérons l'ensemble $\mathcal{C} := \{0, 1\}^Q$ de toutes les combinaisons possibles d'hypothèses nulles et alternatives parmi les $Q$. Pour une configuration donnée $c := (c_1, ..., c_Q) \in \mathcal{C}$, l'hypothèse conjointe $\mathcal{H}^c$ se définit comme suit

$$\mathcal{H}^c := \left( \bigcap_{q:c_q=0} H_0^q \right) \cap \left( \bigcap_{q:c_q=1} H_1^q \right)$$

Etant donné deux sous-ensembles complémentaires $\mathcal{C}_0$ et $\mathcal{C}_1$ de $\mathcal{C}$ nous définissons les hypothèses composites nulle $\mathcal{H}_0$ et alternative $\mathcal{H}_1$ telles que :

$$\mathcal{H}_0 := \bigcup_{c \in \mathcal{C}_0} \mathcal{H}^c, \quad \mathcal{H}_1 := \bigcup_{c \in \mathcal{C}_1} \mathcal{H}^c$$

L'objectif est ici de tester $\mathcal{H}_0$ par rapport à $\mathcal{H}_1$ pour chaque élément $i$ ($1 \leq i \leq n$).

# 3 Modèle

On désigne par $P_i^q$ la probabilité critique obtenue pour le test $q$ sur l'élément $i$. Et on définit le $z$-score : $Z_i^q = -\Phi^{-1}(P_i^q)$, où $\Phi$ représente la fonction de répartition de la loi Gaussienne

standard. On note $Z_i := (Z_i^1, ..., Z_i^Q)$ le vecteur contenant les $z$-scores de l'élément $i$.

A chaque élément $i$ est associé un vecteur $L_i := (L_i^1, ..., L_i^Q) \in \mathcal{C}$, où $L_i^q$ est la variable binaire étant égale à 0 si $H_{0i}^q$ est vraie et 1 si $H_{1i}^q$ est vraie. En supposant que les éléments sont indépendants, chaque vecteur de $z$-scores est issu d'un modèle de mélange avec $2^Q$ composantes défini comme suit :

$$Z_i \sim \sum_{c \in \mathcal{C}} w_c \psi^c. \tag{1}$$

où $\psi^c$ est la loi de $Z_i$ conditionnellement à $L_i = c$ et $w_c = \Pr\{L_i = c\}$.

Le modèle de mélange (1) implique $2^Q$ distributions multivariées $\psi^c$ devant être estimées. Dans la suite, nous faisons l'hypothèse que les fonctions de répartition $\Psi_c$ associées ont la forme suivante :

$$\Psi_c^\theta(Z_i) = C_\theta(F_{c_1}^1(Z_i^1), ..., F_{c_q}^q(Z_i^q), ..., F_{c_Q}^Q(Z_i^Q))$$

où $F_0^q$ (resp. $F_1^q$) est la fonction de répartition marginale de $Z_i^q$ conditionnellement à $L_i^q = 0$ (resp. $L_i^q = 1$) et $C_\theta$ est une fonction copule de paramètre $\theta$ modélisant la structure de dépendance entre les $Q$ z-scores. Les distributions $\psi_c$ s'écrivent alors:

$$\psi_\theta^c(Z_i) = c_\theta \left( F_{c_1}^1(Z_i^1), ..., F_{c_Q}^Q(Z_i^Q) \right) \prod_{q:c_q=0} f_0^q(Z_i^q) \prod_{q:c_q=1} f_1^q(Z_i^q) \tag{2}$$

où $f_0^q$ (resp. $f_1^q$) est la fonction densité marginale de $Z_i^q$ conditionnellement à $L_i^q = 0$ (resp. $L_i^q = 1$). Ainsi, seules $2Q$ fonctions de répartition univariées $F_0^1, ...., F_0^Q, F_1^1, ....F_1^Q$, les probabilités $w_c$ ainsi que le paramètre $\theta$ sont à estimer. Nous considérons ici la copule gaussienne, qui permet de spécifier des corrélations différentes pour chaque paire $(q, q')$ de conditions.

Les paramètres du modèle sont les distributions $f_1^1, ...f_1^Q$, les proportions du mélange $w_c$ et le paramètre de la copule $\theta$. Pour réduire le temps de calcul de la procédure d'inférence, nous proposons de la diviser en deux étapes :

1. Ajuster un modèle de mélange sur chaque ensemble de $z$-scores $\{Z_i^q\}_{1 \leq i \leq n}$ afin d'obtenir une estimation de chaque distribution alternative $f_1^q$.

2. Estimer les proportions $w_c$ de chaque configuration $c$ et le paramètre de la copule $\theta$ à l'aide d'un algorithme EM, après avoir incorporé les estimations $\hat{f}_1^q$.

La distribution marginale des z-scores $Z_i^q$ associés au $q$-ième test peut être déduite du modèle (1) combiné à la définition du $\psi_c$ (2). On a

$$Z_i^q \sim \pi_0^q f_0^q + (1 - \pi_0^q) f_1^q, \tag{3}$$

où $f_0^q$ est la fonction densité de $Z_i^q$ conditionnellement à $L_i^q = 0$ et $f_1^q$ sa fonction densité conditionnellement à $L_i^q = 1$.

Par définition des $z$-scores $Z_i^q$, leur distribution nulle (c'est-à-dire la distribution conditionnelle à $L_i^q = 0$) est la loi Gaussienne standard. On a ainsi $f_0^q = \phi$ pour tous les $q$, où

$\phi$ représente la fonction de densité gaussienne standard. Pour estimer les proportions nulles $\pi_0^q$, utilisons l'estimateur introduite par Storey [15]. Les distributions alternatives $f_1^q$ sont estimées de manière non paramétrique à l'aide d'une méthode d'estimation à noyau.

Les estimations de $\hat{f}_1^q$ peuvent être utilisées pour calculer le $\hat{F}_1^q$. Ces estimations peuvent être directement introduites dans le modéle de mélange, de sorte que les seules quantités à estimer sont les proportions $w_c$ des $2^Q$ composantes de mélange et le paramètre de la copule $\theta$. L'inférence peut être réalisée efficacement à l'aide d'un EM standard.

Notre méthode permet d'obtenir des estimations des probabilités *a posteriori* d'appartenir à une configuration $c \in \mathcal{C}_1$, pouvant être utiliser comme statistique de test, de la manière suivante :

$$\hat{\tau}_i = \sum_{c \in \mathcal{C}_1} \widehat{\Pr}\{L_i = c | Z_i; \hat{\theta}\},$$

# 4 Illustration : Identification de gènes pléiotropes dans une étude portant sur 14 troubles psychiatriques

Afin d'illustrer notre méthode, nous avons analysé conjointement 14 troubles psychiatriques à partir de résultats d'études d'association génétique issues du Psychiatric Genomics Consortium (PGC). Les troubles étudiés sont l'anorexie mentale [4], l'anxiété [10], l'autisme [6], les troubles liés à la consommation d'alcool [12], le trouble obsessionnel compulsif [1], la bipolarité [14], la schizophrénie [7], le stress post-traumatique [9], le syndrome de la Tourette [18], la consommation de cannabis [11], la dépression majeure [17] et le trouble déficitaire de l'attention avec ou sans hyperactivité [3]. Les études d'association génétiques ont été réalisées sur $n = 17,425$ gènes, et l'objectif de l'analyse est d'identifier des gènes pléiotropes, c'est-à-dire associés simultanément à plusieurs troubles. Ici, nous nous sommes intéressés particulièrement aux gènes associés à au moins 8 troubles différents. L'hypothèse composite nulle correspondante pour le gène $i$ est alors :

$$\mathcal{H}_{0i} : \{\text{Le gène } i \text{ est associé à au plus 7 troubles.}\}$$

Avec un seuil $\alpha = 0.05$, notre procédure a identifé 17 gènes associés à au moins 8 troubles psychiatriques, et suggère la présence de 2 "hubs" de gènes très impliqués dans les troubles psychiatriques sur les chromosomes 3 et 17 (Figure 1).

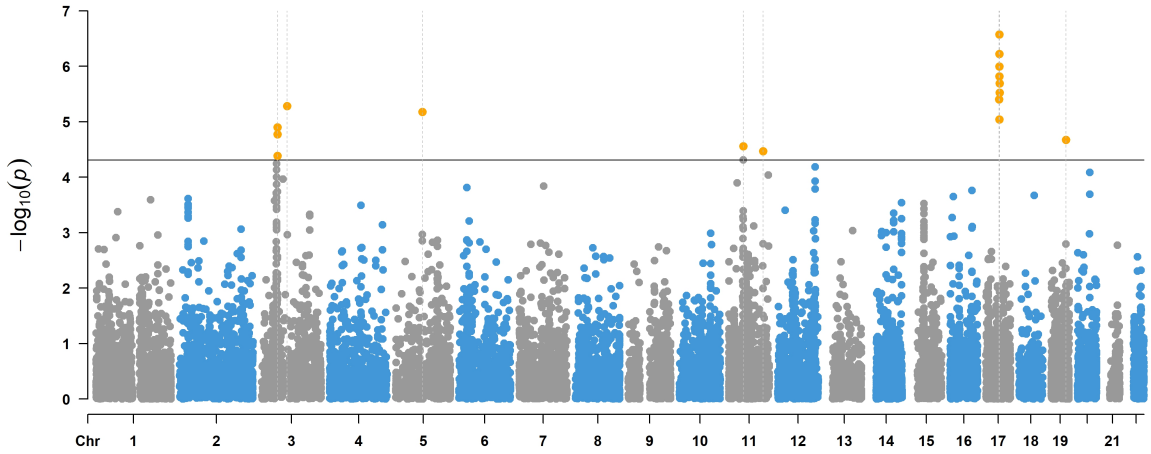



Figure 1: -log10(p-values) du test d'hypothèse composite le long des chromosomes. Les gènes significatifs sont coloriés en orange.

# Bibliographie

# References

[1] Paul D. Arnold et al. "Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis". In: *Molecular Psychiatry* 23 (5 May 2018), pp. 1181–1181.

[2] Roger L Berger. "Multiparameter Hypothesis Testing and Acceptance Sampling". In: 24 (4 1982), pp. 295–300.

[3] Ditte Demontis et al. "Discovery of the first genome-wide significant risk loci for attention-deficit/hyperactivity disorder". In: *Nature genetics* 51 (1 Jan. 2019), p. 63.

[4] Laramie Duncan et al. "Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa". In: *American Journal of Psychiatry* 174 (9 Sept. 2017), pp. 850–858.

[5] Kevin J. Gleason et al. "Primo: Integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits". In: *Genome Biology* 21 (1 Sept. 2020), pp. 1–24.

[6] Jakob Grove et al. "Identification of common genetic risk variants for autism spectrum disorder". In: *Nature genetics* 51 (3 Mar. 2019), p. 431.

[7] Max Lam et al. "Comparative genetic architectures of schizophrenia in East Asian and European populations". In: *Nature genetics* 51 (12 Dec. 2019), pp. 1670–1678.

[8] T. Mary-Huard et al. "Querying multiple sets of $p$-values". In: (2021).

[9] Caroline M. Nievergelt et al. “International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci”. In: *Nature Communications 2019 10:1* 10 (1 Oct. 2019), pp. 1–16.

[10] T. Otowa et al. “Meta-analysis of genome-wide association studies of anxiety disorders”. In: *Molecular psychiatry* 21 (10 Oct. 2016), p. 1391.

[11] Joëlle A. Pasman et al. “GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia”. In: *Nature neuroscience* 21 (9 Sept. 2018), p. 1161.

[12] Sandra Sanchez-Roige et al. “Genome-wide association study meta-analysis of the alcohol use disorders identification test (AUDIT) in two population-based cohorts”. In: *American Journal of Psychiatry* 176 (2 Feb. 2019), pp. 107–118.

[13] Michael E Sobel. “Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models”. In: *Source: Sociological Methodology* 13 (1982), pp. 290–312.

[14] Eli A. Stahl et al. “Genome-wide association study identifies 30 Loci Associated with Bipolar Disorder”. In: *Nature genetics* 51 (5 May 2019), p. 793.

[15] John D. Storey. “A Direct Approach to False Discovery Rates”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64 (3 Aug. 2002), pp. 479–498.

[16] Ziqiao Wang and Peng Wei. “IMIX: a multivariate mixture model approach to association analysis through multi-omics data integration”. In: *Bioinformatics* 36 (22-23 Apr. 2021), pp. 5439–5447.

[17] Naomi R. Wray et al. “Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression”. In: *Nature genetics* 50 (5 May 2018), p. 668.

[18] Dongmei Yu et al. “Interrogating the genetic determinants of Tourette’s syndrome and other tiC disorders through genome-wide association studies”. In: *American Journal of Psychiatry* 176 (3 Mar. 2019), pp. 217–227.

# Inference rapide dans les modèles GLM à copule avec variables explicatives catégorielles en utilisant une procédure IFM -OSCFE

Alexandre Brouste[1], Christophe Dutang[2], Lilit Hovsepyan[1] & Tom Rohmer[3]

[1] *Laboratoire Manceau de Mathématiques, Le Mans Université, F-72000 Le Mans*
[2] *Université Grenoble Alpes, CNRS, Grenoble INP, LJK, F-38000 Grenoble*
[3] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan*

**Résumé.** Dans les modèles linéaires généralisés multivariés à copule, des approches d'estimation basées sur le maximum de vraisemblance (MLE) joint peuvent être coûteuses en temps de calcul. Des méthodes alternatives (IFM) basées sur l'estimation des modèles marginaux là encore par MLE ont été proposés dans la littérature, pouvant là encore se révéler toujours coûteuses malgré le gain évident par rapport au MLE. Dans ce papier nous proposons une approche basée sur l'estimation des modèles marginaux utilisant un estimateur explicite consistant et asymptotiquement efficace proposé dans un papier récent, lorsque toutes les covariables du modèle sont catégorielles. Ce nouvel estimateur permet un gain réel en temps de calcul, sans perte sur la qualité d'estimation des paramètres du modèle en comparaison avec l'approche IFM classique.

**Mots-clés.** GLM, copules, IFM

**Abstract.** In copula multivariate generalized linear models, the approach based on joint maximum likelihood estimator (MLE) may be time consuming. Alternative methods based on inference on the marginals (IFM) which consider MLE marginal estimations was been proposed in the literature. Nevertheless, despite the gain in term of calculation time, these approaches may be again time consuming due to high numbers of explanatory variables or modalities. In this paper, we propose a IFM approach based on an explicit, consistent and asymptotically efficient estimator for the margins which is considered in a recent article when all the explicative variables are categorical. This new estimator allows a real gain in term of computation time comparatively to the classical IFM approach.

**Keywords.** GLM, copulas, IFM

## 1 Modèle d'inférence

Considérons un échantillon $\boldsymbol{Y} = (\underline{\boldsymbol{Y}}_1, \ldots, \underline{\boldsymbol{Y}}_n)$ composé de $\mathbb{R}^s$ vecteurs aléatoires indépendants avec pour $i = 1, \ldots, n$, $\underline{\boldsymbol{Y}}_i = (Y_{i,1}, \ldots, Y_{i,s})$. Les distributions marginales des $Y_{i,j}$, $j = 1, \ldots, s$ sont supposées appartenir à la famille exponentielle de paramètre naturelle $\lambda_{1j}, \ldots, \lambda_{nj}$ à valeur dans un espace $\Lambda_j \subset \mathbb{R}$.

En particulier, la vraisemblance $\mathcal{L}_{ij}$ associée à l'expérience statistique engendrée par $Y_{i,j}$, $i \in 1, \ldots, n$ et $j = 1, \ldots, s$ vérifie

$$\log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j \,|\, y_{i,j}) = \frac{\lambda_{ij}(\boldsymbol{\beta}_j) y_{i,j} - b_j\left(\lambda_{ij}(\boldsymbol{\beta}_j)\right)}{a_j(\phi_j)} + c_j(y_{i,j}, \phi_j), \quad y_{i,j} \in \mathbb{Y} \subset \mathbb{R}, \tag{1}$$

et $-\infty$ si $y_{i,j} \notin \mathbb{Y}$, où $a_j : \mathbb{R} \to \mathbb{R}$, $b_j : \Lambda_j \to \mathbb{R}$ et $c_j : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ sont des fonctions mesurables (supposées connues) et $\phi_j$ est le paramètre de dispersion de la distribution, e.g. McCullagh & Nelder (1989, Section 2.2). Les paramètres $\lambda_{1j}, \ldots, \lambda_{nj}$ dépendent des paramètres auxiliaires inconnus $\boldsymbol{\beta}_j \in B_j \subset \mathbb{R}^{p_j}$, à estimer.

En utilisant une fonction dite de lien $g_j$ deux fois continue différentiable et bijective de $b_j'(\Lambda_j)$ à $\mathbb{R}$, les GLM sont définis par une relation liant l'espérance des observations $\mathbf{E}Y_{i,j}$ aux predicteurs linéaires

$$g_j(\mathbf{E}Y_{i,j}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j = \eta_{ij}, \quad \text{pour tout } \boldsymbol{\beta}_j \in B_j,$$

où $\eta_{ij}$ sont les prédicteurs linéaires, $\boldsymbol{x}_{ij} = (x_{ij}^{(1)}, \ldots, x_{ij}^{(m_j)})$, avec $x_{ij}^{(1)} = 1$ sont des vecteurs constitués par les $m_j$ variables explicatives déterministes. En d'autres termes, les paramètres naturels s'écrivent $\lambda_{ij}(\boldsymbol{\beta}_j) = (b_j')^{-1} \circ g_j^{-1}(\eta_{ij})$.

Dans cette communication, on s'intéresse au cas où pour $j = 1, \ldots, s$, les $m_j$ variables explicatives sont catégorielles avec $d_{\ell,j}$ modalités, $\ell = 1, \ldots, m_j$ et sont encodées en utilisant des variables binaires $x_i^{(\ell),k,j}$ valant 1 si la modalité $k$ de la variable $\ell$ associée à la variable réponse $j$ est choisie, et 0 sinon; voir Brouste et al. (2019, 2022). Sans perte de généralité, nous supposons que $\boldsymbol{x}_{ij}$ et $\boldsymbol{\beta}_j = (\beta_{1,j}, (\beta_{k,j}^{(\ell)})_{k,\ell})$ sont tels que le modèle soit identifiable voir Brouste et al. (2023). De plus, on s'intéresse (pour simplifier l'écriture) au modèle à effets simples uniquement que l'on peut réécrire

$$\begin{cases} g_1\left(\mathbf{E}Y_{i,1}\right) &= \beta_{1,1} + \displaystyle\sum_{\ell=2}^{m_1+1} \sum_{k=1}^{d_{\ell,1}} x_i^{(\ell),k,1} \beta_{k,1}^{(\ell)}, \\ &\vdots \\ g_s\left(\mathbf{E}Y_{i,s}\right) &= \beta_{1,s} + \displaystyle\sum_{\ell=2}^{m_s+1} \sum_{k=1}^{d_{\ell,s}} x_i^{(\ell),k,s} \beta_{k,s}^{(\ell)}. \end{cases}$$

Pour les modèles marginaux, Brouste et al. (2023) ont proposé un estimateur explicite, consistant et asymptotiquement efficace, alternatif au MLE (dont l'obtention par des méthodes itératives de descente de gradient peut être couteux en temps de calcul lorsqu'on considère un grand nombre de variables explicatives ou de modalités).

Dans ce contexte-ci, les variables réponses $Y_{i1}, \ldots, Y_{is}$ ne sont pas supposées indépendantes. Plus précisément nous supposons que la distribution jointe de $(Y_{i1}, \ldots, Y_{is})$ est caractérisée par (1) et par une copule paramétrique $C_\theta$, où le paramètre $\theta$ est à estimer.

En utilisant le théorème de Sklar (1959), la log-vraisemblance de $\boldsymbol{y} = (\underline{\boldsymbol{y}}_1, \ldots, \underline{\boldsymbol{y}}_n)$ peut se réécrire:

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\phi}, \theta | \boldsymbol{y}) &= \sum_{i=1}^{n} \log c_\theta(F_{i1}(y_{i,1}|\boldsymbol{\beta}_1, \phi_1), \ldots, F_{is}(y_{i,s}|\boldsymbol{\beta}_s, \phi_s)) + \sum_{i=1}^{n}\sum_{j=1}^{s} \log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j|y_{i,j}). \\ &= (a) + (b), \end{aligned} \tag{2}$$

où $\mathcal{L}_{ij}$ correspond à la vraisemblance associée à $y_{i,j}$ et $c_\theta$ la densitée de copule donnée par

$$c_\theta(u_1, \ldots, u_s) = \frac{\partial^s C_\theta(u_1, \ldots, u_s)}{\partial u_1 \ldots \partial u_s}.$$

Dans ces modèles multivariées, les paramètres du modèle peuvent être estimés par MLE. Néanmoins les méthodes permettant d'accéder au MLE, peuvent là encore se révéler extrêmement coûteuses en temps de calcul. Une approche alternative est d'estimer les paramètres par une méthode d'inférence sur les marges (IFM) voir (Xu 1996, Joe 1997, 2005). Cette méthode consiste à estimer les paramètres marginaux $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s)$ et de dispersion $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_s)$ de telle sorte à maximiser la somme (b) dans (2), c'est-à dire calculer les MLE des modèles marginaux. Puis, le ou les paramètres de copule $\theta$ sont estimés en utilisant les estimations des paramètres marginaux et de dispersion précédents de telle sorte à maximiser la somme (a) dans (2).

Bien que non efficace, on peut montrer que l'estimation résultante des paramètres (joints) possède de bonne propriétés asymptotiques (consistance forte, distribution asymptotique Gaussienne). Néanmoins, en suivant Brouste et al. (2023), les estimations IFM pourront rester très coûteuses en temps de calcul de par les estimations successives par MLE des modèles marginaux.

Dans cette communication, nous proposons une approche IFM-OSCFE, dans laquelle les estimations marginales seront remplacées par les approches one-step explicites (OSCFE), consistantes et asymptotiquement efficaces proposées dans Brouste et al. (2023). Les propriétés asymptotiques concernant l'estimation jointe des paramètres peuvent être facilement démontrées en suivant (Xu 1996).

# 2 Simulations

Dans la table 1, nous avons considéré un modèle bivarié ($s = 2$) avec deux variables explicatives catégorielles ($m_1 = m_2 = 3$) contenant 2 et 3 modalités ($d_{11} = d_{21} = 2$, $d_{12} = d_{22} = 3$) et $n = 10^5$ observations. Les copules considérées sont Clayton, Frank, Gumbel et Normal, voir Nelsen (2007). Les distributions marginales étaient des distributions gammas avec fonction de lien inverse. Les rhos de Spearman associés aux différentes copules étaient 0.4 ou 0.8, entraînant un paramètre de copule différent selon le type de copule. 100 runs de chaque scénario ont été réalisés. Pour l'ensemble des cas considérés, nous avons obtenu des estimations du paramètre de copule ainsi que des variances aussi proches du vrai paramètre, que ce soit en utilisant un MLE ou en utilisant un OSCFE sur les marginales dans l'approche IFM.

| Spearman's $\rho$ | Copula type | Theo. $\theta$ | (**IFM**) Mean $\theta$ | | (**IFM**) Sd $\theta$ ($\times 10^3$) | |
|---|---|---|---|---|---|---|
| | | | **MLE** | **OSCFE** | **MLE** | **OSCFE** |
| 0.4 | Clayton | 0.758 | 0.758 | 0.758 | 7.431 | 7.431 |
| | Frank | 2.610 | 2.613 | 2.613 | 20.86 | 20.83 |
| | Gumbel | 1.382 | 1.382 | 1.382 | 3.821 | 3.823 |
| | Normal | 0.416 | 0.416 | 0.416 | 2.242 | 2.248 |
| 0.8 | Clayton | 3.188 | 3.186 | 3.187 | 18.03 | 18.03 |
| | Frank | 7.902 | 7.900 | 7.902 | 32.98 | 33.01 |
| | Gumbel | 2.582 | 2.580 | 2.582 | 9.249 | 9.243 |
| | Normal | 0.814 | 0.813 | 0.813 | 1.094 | 1.091 |

Table 1: Valeur moyenne et écart-types des estimations du paramètre de copule $\theta$ en utilisant une approche IFM utilisant une estimation des paramètres marginaux par MLE et par OSCFE, pour $n = 10^5$ observations et $B = 100$ simulations pour le modèle Gamma-GLM bivarié avec liens inverses.

Pour 5 paramètres à estimer par marge, les approches IFM-MLE et IFM-OSCFE étaient comparables en temps de calculs mais pour ce petit nombre de paramètres, les deux approches sont déjà environ 75 fois plus rapides que l'approche MLE (estimation jointe sur 13 paramètres au total). Reprenant les simulations réalisées dans Brouste et al. (2023), à partir de 20 modalités pour chacune des deux variables explicatives, l'estimation des paramètres marginaux était d'environ 95 fois plus rapides en utilisant l'approche OSCFE que l'approche MLE; on peut donc sans trop de risque en conclure que le IFM-OSCFE sera nettement plus rapide que le IFM-MLE dans le cadre d'un modèle GLM multivarié à copule, ouvrant des perspectives prometteuses en terme d'application rapide et sélection de modèle efficace.

# References


Brouste, A., Dutang, C., Hovsepyan, L. & Rohmer, T. (2023), 'One-step closed-form estimator for generalized linear model with categorical explanatory variables', *Statistics and Computing* **33**(6), 138.

Brouste, A., Dutang, C. & Rohmer, T. (2019), 'Closed-form maximum likelihood estimator for generalized linear models in the case of categorical explanatory variables: application to insurance loss modeling', *Computational Statistics* .

Brouste, A., Dutang, C. & Rohmer, T. (2022), 'A closed-form alternative estimator for glm with categorical explanatory variables', *Communications in Statistics-Simulation and Computation* pp. 1–17.

Joe, H. (1997), *Multivariate Models and Multivariate Dependence Concepts*, CRC press.

Joe, H. (2005), 'Asymptotic efficiency of the two-stage estimation method for copula-based models', *Journal of multivariate Analysis* **94**(2), 401–419.

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear models*, Vol. 37, CRC press.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Sklar, M. (1959), 'Fonctions de repartition an dimensions et leurs marges', *Publ. inst. statist. univ. Paris* **8**, 229–231.

Xu, J. J. (1996), Statistical modelling and inference for multivariate and longitudinal discrete response data, PhD thesis, University of British Columbia.

# Données longitudinales

# Régression quantile pénalisée pour des données longitudinales avec hétéroscédasticité.

Angelo Alcaraz [1], Audrey poterie [2] & Gilles Durrieu [3]

[1] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , angelo.alcaraz@univ-ubs.fr*
[2] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , audrey.poterie@univ-ubs.fr*
[3] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , gilles.durrieu@univ-ubs.fr*

**Résumé.** La présence d'hétéroscédasticité dans les données peut souvent mettre en difficulté un processus de modélisation statistique. Dans ce contexte, les modèles mixtes et modèles pour données longitudinales ont été développés. Dans le contexte des modèles mixtes et des données longitudinales, cet article aborde directement ce problème. Plus précisément, notre objectif dans ce travail est de permettre d'améliorer l'estimation de la dispersion des données dans le cadre de données hétéroscédasticités. Pour ce faire, nous développons un nouvel estimateur quantile basé sur la distribution asymétrique de Laplace, qui explique l'hétéroscédasticité entre différents groupes d'individus. Outre le développement de ce nouveau modèle, ce travail établit aussi les bonnes propriétés asymptotiques de cet estimateur sous des hypothèses minimales sur les données et les vérifient à l'aide de simulations. En utilisant le formalisme permissif de la distribution de Laplace asymétrique, nous démontrons les propriétés asymptotiques d'une classe d'estimateurs définis par un problème d'optimisation généralisé inspiré du maximum de vraisemblance. Une pénalisation Ridge est aussi proposée pour traiter les problèmes de surestimation de la variabilité. Plus généralement, cet article présente un modèle permettant de traiter plus précisément les problèmes d'estimation de volume. La méthode a été implémentée en `R`, l'ensemble des fonctions est disponible sur Github.

**Mots-clés.** Distribution de Laplace asymétrique, modèles linéaires mixtes quantiles, quadrature gaussienne, régression Ridge.

**Abstract.** The presence of heteroscedasticity in data can often throw statistical modeling into disarray. In the context of mixed models and longitudinal data, this paper directly addresses this problem. We develop a quantile estimator based on the asymmetric Laplace distribution, which explains the heteroscedasticity between different groups of data. In addition to developing this new model, our paper establishes the good asymptotic properties of this estimator under minimal assumptions on the data and verifies them using simulations. Instead of improving performance point by point, our model focuses on the correct representation of data dispersion. Using the permissive formalism of the asymmetric Laplace distribution, we demonstrate the asymptotic properties of a class of estimators defined by a generalized optimization problem inspired by maximum likelihood. A Ridge penalization is proposed to address problems of variability overestimation. More generally, this paper presents a model for handling volume estimation problems more accurately. The method has been implemented in `R`, and the full set of functions is available on Github.

**Keywords.** Asymmetric Laplace Distribution, Linear Quantile Mixed Models, Gaussian Quadrature, Ridge Regression.

# 1 Introduction

Dans cet article, nous proposons une théorie formelle pour traiter l'hétéroscédasticité dans un modèle de données de panel. Dans certaines applications, il peut être utile de disposer d'un estimateur précis de la dispersion des données, en particulier pour mesurer l'hétéroscédasticité. C'est ce que tente de faire le modèle présenté dans cet article, en mentionnant explicitement les différences de variabilité entre différents groupes de données. De plus, le formalisme mathématique sous-jacent proposé permet d'ajouter simplement des pénalités, ce qui permet de résoudre les problèmes de surestimation de la variance. En étudiant plusieurs variables indépendantes d'intérêt à l'aide du modèle présenté dans cet article, il est alors possible d'effectuer une estimation de volume au sens large (pour des espaces de dimensions supérieures à 1), en contrôlant la dispersion intra-groupe de chacune de ces variables. Considérons le modèle suivant:

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\epsilon}_n \tag{1}$$

où pour tout $n \geq 1$, $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^t$ est le vecteur des observations, $\boldsymbol{X}_n$ est une matrice connue de dimension $n \times p$ de lignes $\boldsymbol{x}_i^t \in \mathbb{R}^p$, $i = 1, \ldots, n$, $\boldsymbol{\epsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^t$ est un vecteur d'erreur i.i.d et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ défini le vecteur de paramètres inconnus à estimer. On appelle $\tau$-ième quantile de régression ( Koenker and Bassett (1978)), $0 < \tau < 1$, toute solution du problème de minimisation :

$$\boldsymbol{\beta}(\tau) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}\right) \tag{2}$$

ou $\rho_\tau(u) = u(\tau - \mathbb{1}_{u<0})$ avec $\mathbb{1}_{\mathcal{P}}$ qui prends les valeurs 1 ou 0 selon la réalisation ou non de la condition $\mathcal{P}$. La fonction $\rho$ est appelée "fonction de perte du quantile" et est classiquement utilisée pour définir le quantile d'une variable aléatoire. Le développement de la régression quantile pour les données de panel suit en quelque sorte le chemin du développement des modèles linéaires généralisés (Liang and Zeger (1986)). Koenker (2004) s'est intéressé aux données longitudinales, une forme classique de données que l'on peut trouver en médecine, en écologie ou en biologie. Dans ce paradigme, nous tenons compte de la dépendance entre des mesures groupées (par exemple, une mesure médicale sur le même sujet). Ceci a conduit à la formulation d'un modèle de régression quantile à effets fixes, exprimé sous la forme d'une pénalité lasso (least absolute shrinkage and selection operator) (Lamarche (2010)).

Nous examinons ici le développement de la régression quantile basée sur la distribution de Laplace asymétrique (Yu and Zhang (2005)). Plusieurs articles fondamentaux ont établi le lien entre le problème de minimisation de la régression quantile et l'estimateur du maximum de vraisemblance ( Geraci and Bottai (2007)). Cela a permis d'utiliser les algorithmes classiques de résolution de l'estimateur du maximum de vraisemblance pour

déterminer l'estimateur de la régression du quantile. On notera l'utilisation de la théorie bayésienne ( Aghamohammadi and Mohammadi (2017), Ji and Shi (2022)) ou de l'algorithme *Expectation-Maximisation* (Geraci and Bottai (2014), Geraci (2014), Galarza et al. (2017) Tian et al. (2020), Battagliola et al. (2022)). Cette communication suit les traces de ces auteurs, en profitant de la distribution de Laplace asymétrique et de son lien avec la régression par quantile et en l'utilisant sur des données longitudinales avec des effets aléatoires.
Nous présentons dans cet article le modèle et quelques résultats théoriques de nos estimateurs, avant d'illustrer les performances de notre approche sur des simulations.

# 2 Modèle et résultats théoriques

## 2.1 Régression quantile pour des données longitudinales hétéroscédastiques utilisant la distribution asymétrique de Laplace

Nous considérons le cadre des données longitudinales. Soit $n \in \mathbb{N}$ le nombre d'individus. L'indice $i \in \{1, \dots, n\}$ dénote le niveau des individus et, pour chaque individu $i$, on considère $n_i \in \mathbb{N}$ mesures. L'indice $j \in \{1, \dots, n_i\}$ correspond à la $j$-ième mesure du $i$-ième individu. Soit $\boldsymbol{Y} = (Y_{ij})_{1 \leq i \leq n, 1 \leq j \leq n_i} \in \mathbb{R}^N$ une variable réponse étudiée, avec $N = \sum_{i=1}^{n} n_i$ . La forme générale du modèle mixte utilisé est la suivante :

$$Y_{ij} = \boldsymbol{X}_{ij}^t \boldsymbol{\beta} + e_i^t \boldsymbol{\nu} + \epsilon_{ij} \tag{3}$$

où pour tous les $i \in \{1, \dots, n\}$ et tous les $j \in \{1, \dots, n_i\}$, $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ sont les observations pour la $j$-ième mesure du $i$-ième individu, $p \in \mathbb{N}$ est le nombre de variables, $\boldsymbol{\beta} \in \mathbb{R}^p$ sont les paramètres de la régression (qui peuvent contenir des effets fixes), $e_i \in \mathbb{R}^n$ est le $i$-ième vecteur de la base standard, $\boldsymbol{\nu} \in \mathbb{R}^n$ est le vecteur des effets aléatoires, décrivant un effet au niveau individuel comme une ordonnée à l'origine, et $\epsilon_{ij}$ est une variable aléatoire centrée représentant le terme d'erreur. Suivant Geraci and Bottai (2014), nous ajoutons deux autres hypothèses sur le modèle. Premièrement, on prend $\boldsymbol{\nu} \sim \mathcal{N}(0, \phi\Sigma)$, avec $\phi \in \mathbb{R}$ un paramètre à fixé et $\Sigma \in \mathbb{R}^{n^2}$ une matrice définie positive connue. Cette dernière est estimée plus loin et représentera la structure de dépendance entre les individus. Deuxièmement, nous supposons que $\epsilon_{ij}$ suit une distribution de Laplace asymétrique, comme dans Geraci and Bottai (2007). Une variable aléatoire $Z$ suit une distribution de Laplace asymétrique de paramètres $(\mu, \sigma, \tau)$ ( qui est dénotée par $Z \sim ALD(\mu, \sigma, \tau)$ ) si sa densité peut être exprimée comme :

$$f_Z(z) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\rho_\tau\left(\frac{z-\mu}{\sigma}\right)\right)$$

avec $\rho_\tau(u) = u(\tau - \mathbb{1}_{u<0})$ désignant la fonction de perte quantile. Une étude détaillée de la distribution de Laplace asymétrique est proposée par Yu and Zhang (2005). Plus précisément, nous supposons que $\epsilon_{ij} \sim ALD(0, \sigma_i, \tau)$, où $0 < \tau < 1$ représente l'ordre du quantile et $\boldsymbol{\sigma} = (\sigma_i)_{1 \leq i \leq n} \in (\mathbb{R}_+^*)^n$ correspond aux paramètres d'échelle. En utilisant cette hypothèse, nous prenons explicitement en compte l'hétéroscédasticité qui existe entre les individus. De plus, nous supposons que tous les $\epsilon_{ij}$ sont indépendants les uns des autres et des effets

aléatoires. Finalement, nous devons déduire le vecteur $\boldsymbol{\theta} = (\boldsymbol{\sigma}, \boldsymbol{\beta}) \in (\mathbb{R}_+^*)^n \times \mathbb{R}^p$. En ce qui concerne l'asymptotique, l'une des conséquences de ce choix de modélisation est la manière de tendre vers l'infini. Comme le nombre de paramètres à estimer dépend maintenant de $n$, nous ne nous intéressons pas à la limite lorsque $n$ tend vers l'infini. Nous déterminons plutôt la limite de notre estimateur lorsque pour tout $1 \leq i \leq n, n_i \to +\infty$, c'est-à-dire lorsque le nombre de mesures par individu tend vers l'infini.
Nous définissons notre estimateur $\hat{\boldsymbol{\theta}}$ comme un estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$. Soit

$$f_{Y_{ij}|\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\sigma}}(y, \boldsymbol{v}) = \frac{\tau(1-\tau)}{\sigma_i} \exp\left(-\rho_\tau\left(\frac{y-\mu_{ij}}{\sigma_i}\right)\right)$$

la densité conditionelle de la réponse par rapport au paramètres et $\mu_{ij} = \boldsymbol{X}_{ij}^t\boldsymbol{\beta} + \boldsymbol{e}_i^t\boldsymbol{v}$. Grâce à l'hypothèse d'indépendance on obtient pour tout $y$ et $\boldsymbol{v}$ la densité multivariée

$$f_{\boldsymbol{Y},\boldsymbol{\nu}|\boldsymbol{\theta}}(y, \boldsymbol{v}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} f_{Y_{ij}|\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\sigma}}(y, \boldsymbol{v}) f_{\boldsymbol{\nu}|\boldsymbol{\phi}}(\boldsymbol{v}). \tag{4}$$

En intégrant (4) sur les effets aléatoires, on obtient la vraisemblance. On a d'abord :

$$f_{\boldsymbol{Y}}(y) = \int_{\mathbb{R}^n} f_{\boldsymbol{Y},\boldsymbol{\nu}|\boldsymbol{\theta}}(y, \boldsymbol{v}) d\boldsymbol{v}. \tag{5}$$

Il en découle que l'estimateur $\hat{\boldsymbol{\theta}}$ est défini par

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\left(\log\left(\int_{\mathbb{R}^n} f_{\boldsymbol{Y},\boldsymbol{\nu}|\boldsymbol{\theta}}(y, \boldsymbol{v}) d\boldsymbol{v}\right)\right) \tag{6}$$

avec $\Theta \subset (\mathbb{R}_+^*)^n \times \mathbb{R}^p$ l'espace des paramètres. Le problème de maximisation est équivalent au problème de minimisation de l'estimateur de la régression quantile (Koenker and Bassett (1978)). Cette équivalence est la principale motivation de l'utilisation de la distribution de Laplace asymétrique. En outre, nous définissons un autre estimateur $\hat{\boldsymbol{\theta}}_f$ de $\boldsymbol{\theta}$ qui est défini comme la solution du problème de maximisation suivant

$$\hat{\boldsymbol{\theta}}_f = \arg\max_{\boldsymbol{\theta}\in\Theta} \int_{\mathbb{R}^n} \left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_\tau\left(\frac{y-\mu_{ij}}{\sigma_i}\right) + \sum_{i=1}^{n} n_i \ln(\sigma_i) + \sum_{i=1}^{n} f_{n_i}(\boldsymbol{\theta})\right) dv \tag{7}$$

où $(f_{n_i})_{1\leq i\leq n}$ est une suite de fonctions prenant ses valeurs dans l'espace des paramètres $\Theta$. Ce problème d'optimisation correspond au problème d'optimisation du maximum de vraisemblance avec un terme de pénalité $\sum_{i=1}^{n} f_{n_i}(\boldsymbol{\theta})$ supplémentaire. Ce terme de régularisation peut être ajouté au problème d'optimisation de maximum de vraisemblance dans le but par exemple de réduire la complexité du modèle et d'éviter l'*overfitting* (Tibshirani (1996), Friedman et al. (2000)).

## 2.2 Résultats théoriques

Afin d'étudier le comportement asymptotique de nos estimations, il est nécessaire d'introduire plusieurs hypothèses.

***A1***. On a $N \to +\infty$ et pour tout $i \in \{1, \dots, n\}$, $n_i \to +\infty$ tel que $n_i = \mathcal{O}(N)$.

***A2***. La vraie valeur de $\boldsymbol{\theta}$ est notée $\boldsymbol{\theta}^0$ et on a $\boldsymbol{\theta}^0 \in \text{int}(\Theta)$, où $\text{int}(\Theta)$ est l'intérieur de $\Theta$.

***A3***. Pour tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, n_i\}$, les variables aléatoires $\epsilon_{ij}$ sont mutuellement indépendantes et indépendantes de $\nu$.

***A4***. La suite $\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ tends vers $c_i \in \mathbb{R}^p$ où $n_i \to +\infty$ pour tout $i \in \{1, \dots, n\}$.

***A5***. La suite $\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} X_{ij}^t$ tend vers une matrice définie positive $C_i \in \mathbb{R}^{p \times p}$ où $n_i \to +\infty$ for all $i \in \{1, \dots, n\}$.

***A6***. Pour tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, n_i\}$, les variables aléatoires $\mathbb{E}[\epsilon_{ij} | \boldsymbol{Y}]$ sont mutuellement indépendantes et indépendantes de $\mathbb{E}[\nu | \boldsymbol{Y}]$.

L'hypothèse (***A1***) permet de définir la façon dont notre modèle va atteindre l'infini, comme explicité dans la partie 2.1. L'hypothèse (***A2***) est une hypothèse classique pour la régression quantile et est nécessaire à l'application du théorème principal utilisé lors de la preuve du comportement asymptotique de notre estimateur. Les hypothèses (***A3***) à (***A5***) limitent la forme que peuvent prendre les données et garantissent l'existence de moments du premier et du second ordre. Ces hypothèses sont également couramment utilisées dans la régression quantile. Nous pouvons constater que ni les données ni le terme de variabilité $\epsilon_{ij}$ n'ont besoin de suivre une distribution normale pour que les résultats asymptotiques soient établis. La dernière hypothèse (***A6***) est nécessaire pour la preuve du Théorème 1 et est rarement utilisée dans la littérature. On peut voir une utilisation et une explication détaillée de cette hypothèse dans Weidenhammer (2017). Les principaux résultats de ce document sont les théorèmes suivants qui établissent la normalité asymptotique de nos estimateurs.

**Théorème 1** *Sous les hypothèses (**A1**) à (**A6**), on a, lorsque $N$ tends vers l'infini, la normalité asymptotique suivante*

$$diag((\sqrt{N}, \sqrt{N}))\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \begin{pmatrix} \sqrt{N}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^0) \\ \sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, B^{-1}(\boldsymbol{\theta}^0)) \tag{8}$$

*où $B(\theta)$ est une matrice définie positive de taille $(n+p) \times (n+p)$.*

De plus, nous pouvons calculer la valeur de $B(\boldsymbol{\theta}^0)$, ce qui nous permet d'établir des intervalles de confiance précis. Cela nous permet également de comparer la distribution asymptotique du paramètre avec celle de l'estimateur et nous renseigne sur la vitesse de convergence en $\sqrt{N}$. La cohérence de notre estimateur découlera immédiatement de ce résultat, comme dans Weidenhammer (2017). Nous définissons deux autres hypothèses sur $(f_{n_i})_{1 \leq i \leq n}$ de l'équation (7) :

***C1.*** Pour tout $1 \leq i \leq n$, on a $\frac{f_{n_i}}{n_i} \longrightarrow l \in \mathbb{R}$ lorsque $n_i$ tends vers l'infini.

***C2.*** L'application

$$\begin{aligned}\Phi : \quad \Theta &\longrightarrow (\mathbb{R}_+^*)^n \\ \boldsymbol{\theta} &\longrightarrow \left(\sigma_i \exp\left(\frac{f_{n_i}(\boldsymbol{\theta})}{n_i}\right)\right)_{1\leq i\leq n}\end{aligned}$$

est une fonction bijective.

**Théorème 2** *Sous les hypothèses (**A1**) à (**A6**), (**C1**) et (**C2**), on a, lorsque $N$ tends vers l'infini, la normalité asymptotique suivante*

$$diag((\sqrt{N},\sqrt{N}))\ (\hat{\boldsymbol{\theta}}_f - \tilde{\boldsymbol{\theta}}^0) = \begin{pmatrix}\sqrt{N}(\hat{\boldsymbol{\sigma}}_f - \tilde{\boldsymbol{\sigma}}^0)\\ \sqrt{N}(\hat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}^0)\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tilde{B}^{-1}(\boldsymbol{\theta}^0)) \tag{9}$$

*où $\tilde{\boldsymbol{\theta}}^0 = (\tilde{\boldsymbol{\sigma}}^0, \boldsymbol{\beta}^0)$, pour tout $i \leq n$, $\tilde{\sigma}_i^0 = \exp\left(\frac{f_{n_i}(\boldsymbol{\theta}^0)}{n_i}\right)\sigma_i^0$ et*

- $\tilde{B}_{\boldsymbol{\sigma},\boldsymbol{\sigma}}(\theta^0) = e^{2l} B_{\boldsymbol{\sigma},\boldsymbol{\sigma}}(\theta^0)$,
- $\tilde{B}_{\boldsymbol{\sigma},\boldsymbol{\beta}}(\theta^0) = e^{l} B_{\boldsymbol{\sigma},\boldsymbol{\beta}}(\theta^0)$,
- $\tilde{B}_{\boldsymbol{\beta},\boldsymbol{\beta}}(\theta^0) = B_{\boldsymbol{\beta},\boldsymbol{\beta}}(\theta^0)$.

Le Théorème 2 permet d'assurer la normalité asymptotique pour une toute nouvelle classe d'estimateurs $\boldsymbol{\theta}_f$. Par conséquent, en utilisant une suite appropriée de fonctions $(f_{n_i})_{1\leq i\leq n}$, nous pouvons modifier le problème d'optimisation (7) tout en conservant les mêmes garanties théoriques. Ici, nous nous intéressons au cas où la fonction $f$ est une pénalisation de Ridge et nous voulons donc résoudre

$$\hat{\boldsymbol{\theta}}_f = \arg\max_{\boldsymbol{\theta}\in\Theta} \int_{\mathbb{R}^n} \left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_\tau\left(\frac{y-\mu_{ij}}{\sigma_i}\right) + \sum_{i=1}^{n} n_i \ln(\sigma_i) + \lambda \sum_{i=1}^{n} \sigma_i^2\right) dv. \tag{10}$$

Nous pouvons appliquer le Théorème 2 à (10) dans lequel $f_{n_i}(\boldsymbol{\theta}) = \lambda\sigma_i^2$, pour tout $1 \leq i \leq n$ et $\lambda > 0$ est un paramètre fixé. Nous avons la condition (***C1***) car, pour tout $1 \leq i \leq n$, nous avons $\frac{f_{n_i}}{n_i} \longrightarrow 0$. Dans ce cas, l'hypothèse (***C2***) est aussi vérifiée. En effet, si nous écrivons $\tilde{\sigma}_i = \sigma_i \exp(\lambda\frac{\sigma_i^2}{n_i})$, nous pouvons obtenir les paramètres d'échelle désirés $\sigma_i$ de la manière suivante :

$$\sigma_i = \frac{\sqrt{W\left(\frac{2\lambda\tilde{\sigma}_i^2}{n_i}\right) n_i}}{\sqrt{2\lambda}}$$

avec $W$ la fonction de Lambert, définie comme la fonction inverse de $x \to xe^x$.
En conséquence, l'estimateur défini par le problème d'optimisation (7) incluant un terme de pénalisation Ridge $\lambda\sum_{i=1}^{n}\sigma_i^2$ est asymptotiquement normal, avec la même matrice de variance-covariance que le problème non pénalisé (car $l = 0$ ici). L'ajout d'une telle pénalisation pourrait permettre un meilleur contrôle sur l'estimation des paramètres d'échelle du modèle. De plus, si $Z \sim ALD(\mu, \sigma, \tau)$, nous avons le résultat suivant (Yu and Zhang (2005))

$$Var(Z) = \frac{\sigma^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}. \tag{11}$$

Comme $\sigma_i$ est le paramètre d'échelle pour le $i$-ième individu, on peut observer que la pénalisation de $\sigma_i$ nous permet de contrecarrer la surestimation éventuelle de la variance entre les mesures d'un même individu. D'une certaine manière, l'information sur le paramètre d'échelle $\sigma_i$ nous renseigne sur la dispersion ou la diversité des mesures pour un individu.

La méthode d'estimation de notre modèle n'est pas détaillé ici. Suivant Geraci (2014), notre approche s'appuie sur un algorithme d'optimisation itératif et l'utilisation d'une quadrature de Gauss. Une implémentation en `R` est disponible sur Github ici : `https://github.com/I621974/hlqmm`. La prochaine partie décrit quelques résultats de simulations justifiant des bonnes propriétés de nos estimateurs.

# 3 Étude par simulation

Cette section est consacrée à des expériences numériques afin d'évaluer la performance de notre méthode.

Pour chaque cas, nous considérons : $p = 9$, $J = 1000$, $n = 6$, $\tau = 0.5$, $\psi = 2$, $\boldsymbol{\sigma}^0 = (1, 2, 1, 3, 1, 0.5)^t$, $\boldsymbol{\beta}^0 = (\psi, 1, 2, -2, -1, 3, 5.5, 0.5, 0.1)^t$, $\phi = 5$ et pour tout $i \in \{1, \ldots, n\}$, $n_i = J = 1000$. Le paramètre $\psi$ représente la valeur réelle de l'ordonnée à l'origine. Nous considérons ensuite trois distributions pour générer les covariables $\boldsymbol{X}$, à savoir une distribution normale $\mathcal{N}(2, 2)$, une distribution de Laplace $\mathcal{L}(2, 3)$ et une distribution de Bernoulli $\mathcal{B}(0.33)$. Nous représentons l'histogramme pour $\hat{\boldsymbol{\sigma}}$ et $\hat{\boldsymbol{\beta}}$ de 1000 réplications de Monte Carlo par coordonnées et on le compare à la distribution normale théorique asymptotique. Les Figures 1 et 2 montrent les distributions empiriques de $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\sigma}}$ respectivement lorsque les covariables sont générées selon la distribution normale $\mathcal{N}(2, 2)$.

Tout d'abord, nous pouvons constater que toutes les distributions empiriques correspondent aux distributions théoriques, à l'exception de l'ordonnée à l'origine. Il existe un biais sur l'ordonnée à l'origine, qui a été supprimé dans la Figure 1 en centrant l'estimateur de l'ordonnée à l'origine pour voir l'adéquation en termes de variance. On peut trouver quelques informations concernant la présence de ce biais et la nécessité d'inclure l'ordonnée à l'origine dans une régression quantile dans Jurečková (1984) et Battagliola et al. (2022). Sinon, nous pouvons voir que tous les estimateurs semblent suivent la distribution normale théorique ($p > 0.05$, test de Shapiro). Le comportement pour d'autres distributions de co-variables est en grande partie le même. Les figures relatives à ces distributions sont omises ici.

Nous nous intéressons ensuite à la convergence de notre estimateur $\hat{\theta}$. Nous utilisons le même *design* de simulation, mais nous faisons maintenant varier le nombre de mesures, $J$. Nous pouvons observer la convergence de l'algorithme dans la Figure 3 pour le paramètre $\boldsymbol{\sigma}$ lorsque $J$ augmente. Ici, les résultats sont assez similaires, quelle que soit la distribution utilisée pour générer les covariables. En regardant l'échelle sur l'axe des ordonnées, on constate que plus le paramètre est proche de 0, plus la méthode a du mal à l'estimer correctement. Les résultats sont les mêmes pour le paramètre $\boldsymbol{\beta}$.

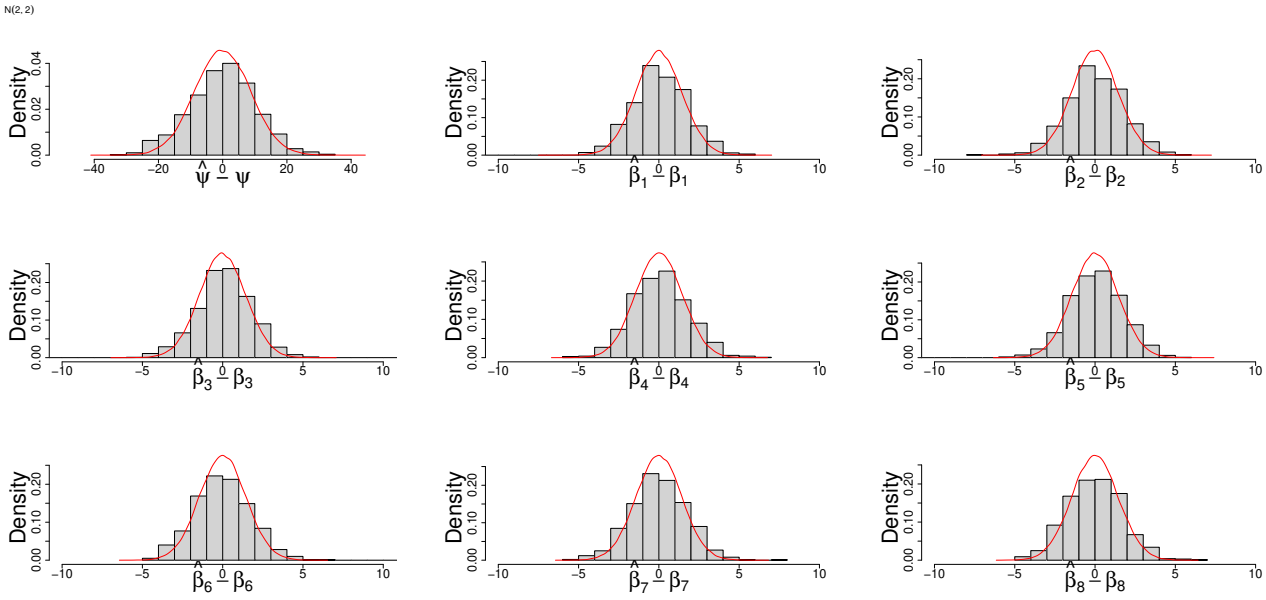



Figure 1: Distribution empirique de $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ pour une génération normale de covariables suivant $\mathcal{N}(2, 2)$. Pour l'estimateur $\hat{\psi}$, le biais empirique a été supprimé.

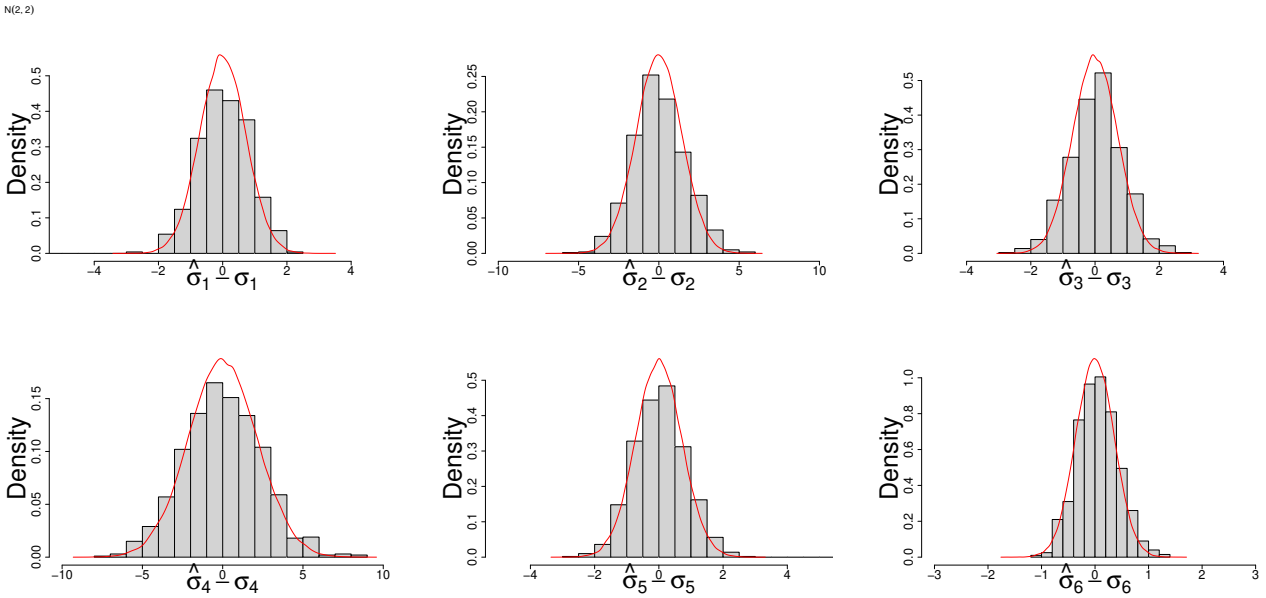



Figure 2: Distribution empirique de $\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}$ pour une génération normale de covariables suivant $\mathcal{N}(2, 2)$.

# 4 Conclusion

Cet article présente une nouvelle classe d'estimateurs basées sur la régression quantile pour modèle mixte qui permettent de préciser et maîtriser l'hétéroscédasticité dans des données

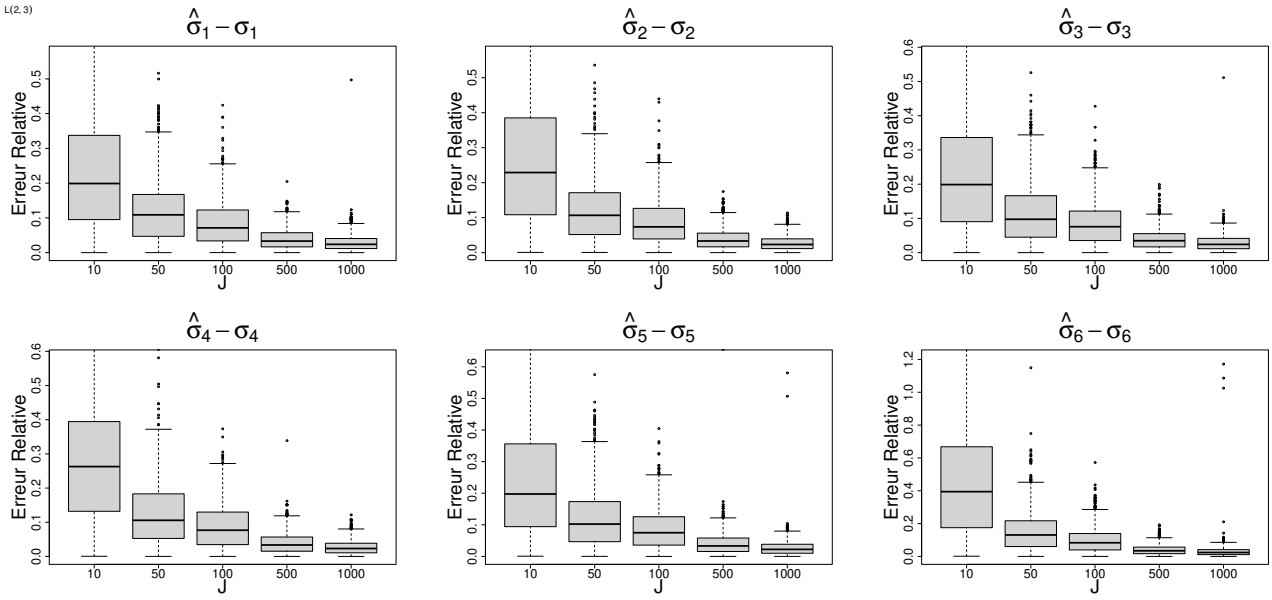



Figure 3: Illustraion de la convergence de $\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^0$ pour une génération de Laplace de covariables suivant $\mathcal{L}(2, 3)$.

longitudinales. Les bonnes propriétés des estimateurs sont montrés théoriquement et soutenues par des simulations. L'intérêt de la méthode sera explicitée via une application écologique.

# Bibliographie


Aghamohammadi, A. and Mohammadi, S. (2017), 'Bayesian analysis of penalized quantile regression for longitudinal data', *Statistical Papers* **58**, 1035–1053.

Battagliola, M. L., Sørensen, H., Tolver, A. and Staicu, A.-M. (2022), 'A bias-adjusted estimator in quantile regression for clustered data', *Econometrics and Statistics* **23**, 165–186.

Friedman, J., Hastie, T. and Tibshirani, R. (2000), 'Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)', *The Annals of Statistics* **28**(2), 337 – 407.

Galarza, C. E., Lachos, V. H. and Bandyopadhyay, D. (2017), 'Quantile regression in linear mixed models: a stochastic approximation em approach', *Statistics and its Interface* **10**(3), 471.

Geraci, M. (2014), 'Linear quantile mixed models: the lqmm package for laplace quantile regression', *Journal of Statistical Software* **57**, 1–29.

Geraci, M. and Bottai, M. (2007), 'Quantile regression for longitudinal data using the asymmetric laplace distribution', *Biostatistics* **8**(1), 140–154.

Geraci, M. and Bottai, M. (2014), 'Linear quantile mixed models', *Statistics and computing* **24**, 461–479.

Ji, Y. and Shi, H. (2022), 'Shrinkage estimation of fixed and random effects in linear quantile mixed models', *Journal of Applied Statistics* **49**(14), 3693–3716.

Jurečková, J. (1984), 'Regression quantiles and trimmed least squares estimator under a general design', *Kybernetika* **20**(5), 345–357.

Koenker, R. (2004), 'Quantile regression for longitudinal data', *Journal of multivariate analysis* **91**(1), 74–89.

Koenker, R. and Bassett, J. G. (1978), 'Regression quantiles', *Econometrica: journal of the Econometric Society* pp. 33–50.

Lamarche, C. (2010), 'Robust penalized quantile regression estimation for panel data', *Journal of Econometrics* **157**(2), 396–408.

Liang, K.-Y. and Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

Tian, Y., Wang, L., Tang, M., Zang, Y. and Tian, M. (2020), 'Likelihood-based quantile autoregressive distributed lag models and its applications', *Journal of Applied Statistics* **47**(1), 117–131.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Weidenhammer, B. (2017), *The Consistency of Quantile Regression in Linear Mixed Models*, Freie Universitaet Berlin (Germany).

Yu, K. and Zhang, J. (2005), 'A three-parameter asymmetric laplace distribution and its extension', *Communications in Statistics—Theory and Methods* **34**(9-10), 1867–1879.

# Clustering Longitudinal Mixed Data

Francesco Amato [1], Julien Jacques [1]

[1] *Univ Lyon, Univ Lyon 2, ERIC, Lyon.*
{*francesco.amato, julien.jacques*}*@univ-lyon2.fr*

**Résumé** Nous présentons un algorithme de clustering pour données longitudinales mixtes. En supposant que les variables non continues sont la discrétisation de variables continues latentes, le modèle s'appuie sur un mélange de lois normales matricielles, capable de prendre en compte simultanément des structures de dépendance entre variables et temporelles. Le modèle est ainsi capable de modéliser simultanément l'hétérogénéité des données, l'association entre les réponses et la structure de dépendance temporelle. Un algorithme EM est développé pour l'estimation des paramètres.
**Mots-clés** Clustering probabiliste. Données longitudinales mixtes. Données à trois voies. Modèles de mélange. Lois Gaussiennes matricielles.

**Abstract.** We present a model-based clustering algorithm to cluster longitudinal mixed data. Assuming that the non-continuous variables are the discretization of underlying latent continuous variables, the model relies on a mixture of matrix-variate normal distributions, accounting simultaneously for within- and between-time dependence structures. The model is thus able to concurrently model the heterogeneity, the association among the responses and the temporal dependence structure. An EM algorithm is developed for parameters estimation.
**Keywords.** Model-based Clustering. Mixed longitudinal data. Three-way data. Mixture models. Matrix-variate Gaussians.

## 1 Context

In many areas of humanities and social sciences, the studies are based on questionnaires completed by participants. Often, these questionnaires are completed several times over the study period. The researchers then analyse these questionnaires to determine typical behaviours within the studied population.
However, the statistical analysis of these questionnaires is far from simple, for several reasons. First, the answers to the questions are often of different types. The analysis of such mixed data is a current research problem in the fields of statistics and machine learning. The second scientific obstacle is the modelling of the temporal evolution of the answers to the questions. Currently, too frequently the analyses are done independently at each temporal phase, then researchers try *a posteriori* to find links between these different analyses, by seeking from one phase to the other to find similar typical behaviour. We can

for example cite Selosse et al., 2019 in the case of clustering of longitudinal ordinal data for an application in psychology. The ideal way to model these data would be through modelling all the responses to the questionnaires at the same time.
In this work we aim at providing a tool to perform model-based clustering on questionnaires repeated over time. Probabilistic (or model-based) clustering offers the advantage of clearly stating the assumptions behind the clustering algorithm, and allows cluster analysis to benefit from the inferential framework of statistics to address some of the practical questions arising when performing clustering (Bouveyron et al., 2019).

## 2 Related work

While several approaches exist for the clustering longitudinal and mixed data separately, literature is poor when they are to be dealt with simultaneously.
An approach to clustering longitudinal data consists in arranging the data in a three-way format and modelling them through a matrix-variate mixture model. This approach offers the advantage of accounting for the overall time-behavior, grouping together the units that have a similar pattern across and within time. While not being new (Basford and McLachlan, 1985), matrix-variate distributions have recently gained attention, and mixtures of matrix-normals (MMN) have been developed and applied both in a frequentist framework in Viroli, 2011a and within a Bayesian one by Viroli, 2011b. These models represent a natural extension of the multivariate normal mixtures to account for temporal (or even spatial) dependencies, and have the advantage of being also relatively easy to estimate by means of EM algorithm (a nice short description of the EM application to MNN is provided in §2.1 of Wang and Melnykov, 2020). More recently, in Gallaugher and McNicholas, 2018 and Melnykov and Zhu, 2018, 2019 extensions for non-normal skewed cases have been proposed and applied. However, matrix-variate models suffer from over-parametrization that leads to estimation issues. To overcome this issue a more parsimonious model (Sarkar et al., 2020) and a new R package (Zhu, Sarkar, and Melnykov, 2022) has been proposed. Despite their efficacy, up to now these methods have only been applied to continuous data.
Our model expands the use of matrix-variate mixtures to mixed data, by building on the framework proposed by McParland and Gormley, 2016 and further developed by Choi, Ahn, and Kim, 2023.

## 3 Preliminaries

Let $Z \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, that is a matrix-variate normal distribution where $M \in \mathbb{R}^{J\times T}$ is the matrix of means, $\Phi \in \mathbb{R}^{T\times T}$ is a covariance matrix containing the variances and covariances between the $T$ occasions or times and $\Sigma \in \mathbb{R}^{J\times J}$ is the covariance matrix

containing the variance and covariances of the $J$ variables. The matrix-normal probability density function (pdf) is given by

$$f(Z|M,\Phi,\Sigma) = (2\pi)^{-\frac{TJ}{2}}|\Phi|^{-\frac{J}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\text{tr}[\Sigma^{-1}(Z-M)\Phi^{-1}(Z-M)^{\intercal}]\right\}. \quad (1)$$

The matrix-normal distribution represents a natural extension of the multivariate normal distribution, since if $Z \sim \mathcal{MN}_{(J\times T)}(M,\Phi,\Sigma)$, then $\text{vec}(Z) \sim \mathcal{MVN}_{JT}(\text{vec}(M),\Phi\otimes\Sigma)$, where vec(.) is the vectorization operator, that is the function mapping from a a $J\times T$ matrix to a $JT$-dimensional vector, and $\otimes$ denotes the Kronecker product. The property of rewriting the general covariance matrix $\Psi\in\mathbb{R}^{JT\times TJ}$ as $\Psi=\Phi\otimes\Sigma$ is called separability condition. Then, the mean and the variance of the multivariate normal normal distribution are:

$$\mathbb{E}(\text{vec}(Z)|M,\Phi,\Sigma) = \text{vec}(M) \quad \text{and} \quad \mathbb{V}(\text{vec}(Z)|M,\Phi,\Sigma) = \Sigma\otimes\Phi. \quad (2)$$

Being a special case of the multivariate normal distribution, the matrix-normal distribution shares the same properties, like, for instance, closure under marginalization, conditioning and linear transformations (Gupta and Nagar, 2000). The separability condition of the covariance matrix has two advantages. First, it allows the modeling of the temporal pattern of interest directly on the covariance matrix $\Phi$. Second, it represents a more parsimonious solution than that of the unrestricted $\Phi\otimes\Sigma$.

Introduced by Viroli, 2011a, the pdf of the finite Mixture of Matrix-Normals (MMN) model is given by

$$f(Z|\boldsymbol{\pi},\boldsymbol{\Theta}) = \sum_{k=1}^{K}\pi_k\phi^{(J\times T)}(Z|M_k,\Phi_k,\Sigma_k),$$

where $\phi^{(J\times T)}$ represents the density function of a $J\times T$-dimensional matrix-variate normal, $K$ is the number of mixture components, $\boldsymbol{\pi}=\{\pi_k\}_{k=1}^K$ is the vector of mixing proportions, subject to constraint $\sum_{k=1}^K\pi_k=1$ and $\boldsymbol{\Theta}=\{\Theta_k\}_{k=1}^K$ is the set of component-specific parameters with $\Theta_k=\{M_k,\Phi_k,\Sigma_k\}$.

# 4 Model

Denote by $y_{ijt}$ the observation of the $j$-th ($j=1,..,J$) variable for the $i$-th ($i=1,\ldots,N$) unit at time $t$ ($t=1,\ldots,T$), that is: imagine to observe $N$ units and measuring $J$ different mixed variables $T$ times throughout the course of the study. We can divide the $J$ mixed variables into $C$ continuous variables and $O$ the non-continuous ones, such that $C+O=J$. Let us reorganize this data in a random-matrix form such that $\mathbf{Y}=\{Y_i\}_{i=1}^N$ is a sample of $J\times T$-variate matrix observations $Y_i=(y_{ijt})\in[\mathbb{R}^{C\times T},\mathbb{N}^{O\times T}]^{\intercal}$, $J=C+O$. The ordered classes are coded by positive integers such that each ordinal variable $o$ the

ordinal levels are $\{1, 2, \ldots, C_o\}$, while the binary classes are coded as 0 and 1.
Assuming our population is heterogeneous and partitioned into $K$ clusters, we define $\ell_i = (\ell_{i1}, \ldots, \ell_{iK})$ as a one-hot encoding representation of group membership, such that $\ell_{ik} = 1$ if the $i$-th unit belongs to the $k$-th cluster.
Then, we can assume that each variable $y_{ijt}$ is the manifestation of an underlying latent continuous variable $z_{ijt}$.

## 4.1 Modelling continuous variables

We assume that the observed continuous variables $y_{ijt}$ match exactly the latent variable:

$$y_{ijt} = z_{ijt}$$

## 4.2 Modelling ordinal variables

To map ordinal data, we follow Amato, Jacques, and Prim-Allaz, 2024. Let the generic ordinal $o$-th variable have $C_o$ levels. Let $\gamma_o$ denote a $C_o + 1$ -dimensional vector of thresholds that partition the real line for the corresponding $o$-th underlying continuous variable, and let the threshold parameters be constrained such that $-\infty = \gamma_{o,0} \leq \gamma_{o,1} \leq \ldots \leq \gamma_{o,C_o} = \infty$. If the latent $z_{i,o,t}$ is such that $\gamma_{o,c-1} < z_{i,o,t} < \gamma_{o,c}$ then the observed ordinal response, $y_{i,o,t} = c$.
Moreover, let define $\mathcal{O}^{O \times T}$ the set of ordinal matrices of size $J \times T$ whose row $o$ takes values in $\{1, \ldots, C_o\}$. Each element of $\mathcal{O}^{O \times T}$ is called a response pattern. Let $R$ be the cardinality of $\mathcal{O}^{O \times T}$. Each response pattern $Y_r \in \mathcal{O}^{O \times T}$ is generated by a portion $\Omega_r$ of the latent space $\mathbb{R}^{O \times T}$ according to thresholds $\boldsymbol{\gamma} := \{\gamma_o\}_{o=1}^{O}$. Let the binary vector $\tilde{Y}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iR})$ be one-hot encoding of $Y_i$ such that if the $r$-th pattern is observed then $\tilde{Y}_{ir} = 1$ and any other entry in the vector equals zero.

A key point is of course the choice of the thresholds $\boldsymbol{\gamma} = \{\gamma_j\}_{o=1}^{O}$. To avoid identifiability and computational complexity issues, thresholds are fixed and not considered as parameters. There are different ways to do it. We decide to follow Corneli, Bouveyron, and Latouche, 2020, where the thresholds are chosen as $\gamma_o = (-\infty, 1.5, 2.5, \ldots, C_o - 0.5, \infty)$ .

## 4.3 Modelling categorical variables

For non-ordered categorical data with $P$ levels we can consider a one-hot encoding for $P - 1$ levels and treat them as binary variables. Binary variables can be considered as a special case of ordinal variables where the number of classes $C_o = 2$. The threshold cutting the underlying continuous variable is set to 0.

### 4.4 Joint model

At this point, we can assume that each observed matrix $Y_i$ is indeed the manifestation of a latent random matrix $Z_i$, and that this underlying random matrix is linked through different relations to the observed matrix $Y_i$, depending on the type of variable each element $y_{i,j,t}$, as described in Section 4.

So, we can think of $Y_i$ as a block matrix, and conveniently split it between the first $C$ rows, representing the observed continuous variables, and the remaining $J - C = O$ rows, representing the observed ordinal and categorical variables. Notice that the slicing happens just over rows but not over columns. Then, for notation's sake we can write $Y_i = [Y_i^\alpha, Y_i^\beta]^\intercal$, where $Y_i^\alpha \in \mathbb{R}^{C\times T}$ is the block containing the continuous variables and $Y_i^\beta \in \mathbb{N}^{O\times T}$ gathers the ordinal and categorical ones (that we coded via integers).

$$\begin{pmatrix}\mathbb{R}^{C\times T}\\ \mathbb{N}^{O\times T}\end{pmatrix} \ni Y_i = \begin{pmatrix} y_{i,1,1} & \cdots & y_{i,1,t} & \cdots & y_{i,1,T}\\ \vdots & \ddots & \vdots & \cdots & \vdots\\ y_{i,j,1} & \cdots & y_{i,j,t} & \cdots & y_{i,j,T}\\ \vdots & \cdots & \vdots & \ddots & \vdots\\ y_{i,J,1} & \cdots & y_{i,J,t} & \cdots & y_{i,J,T}\end{pmatrix} \longleftarrow Z_i = \begin{pmatrix} z_{i,1,1} & \cdots & z_{i,1,t} & \cdots & z_{i,1,T}\\ \vdots & \ddots & \vdots & \cdots & \vdots\\ z_{i,j,1} & \cdots & z_{i,j,t} & \cdots & z_{i,j,T}\\ \vdots & \cdots & \vdots & \ddots & \vdots\\ z_{i,J,1} & \cdots & z_{i,J,t} & \cdots & z_{i,J,T}\end{pmatrix} \in \mathbb{R}^{J\times T}$$

Again, we can write $Z = [Z_i^\alpha, Z_i^\beta]^\intercal$, applying the same logic as for $Y_i$. Then, we assume a mixture of matrix-normal distributions on the latent space $Z_i$.

Assuming that:

$$\ell_i \sim \mathcal{M}(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} := (\pi_1, \dots, \pi_K)$$
$$Z_i|\ell_{ik} = 1 \sim \mathcal{MN}_{(J\times T)}(Z_i|\Theta_k),\ \Theta_k := \{M_k, \Phi_k, \Sigma_k\},$$

we get:

$$f(\ell_i) = \prod_{k=1}^{K} \pi_k^{\ell_{ik}}\ ;\ f(Z_i|\ell_i) = \prod_{k=1}^{K} \left[\phi^{(J\times T)}(Z_i|\Theta_k)\right]^{\ell_{ik}}\ ;$$

where $\mathcal{M}$ indicates the multinomial distribution.

In the following, $\boldsymbol{Z} := \{Z_i\}_{i=1}^N, \boldsymbol{\ell} := \{\ell_i\}_{i=1}^N$ will indicate the ensembles of $Z_i, \ell_i$. Finally, let $\mathbf{Y} := \{Y_i\}_{i=1}^N$ be the collection of the observed matrices $Y_i$.

## 5 Estimation

To estimate the model, since we do not observe neither $Z$ nor $\ell$, we resort to the EM algorithm (Dempster, Laird, and Rubin, 1977).

The EM algorithm is an iterative algorithm alternates two steps: the expectation step (E-step) and the maximization step (M-step). It start from an initialization $\hat{\boldsymbol{\Theta}}^{(0)}$ of the parameters. Then, let denote with the superscript $(s+1)$ the parameters estimated in the current step and with $(s)$ the ones computed in the previous step.
The E-step consists of evaluating $\mathcal{Q}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}^{(s)}) := \mathbb{E}(\log \mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$, that is the expectation of the complete log-likelihood conditioned on the parameters computed in the previous step and on the observed data.
In the M-step the parameters are updated by maximizing the expected complete log-likelihood found on the E step, that is $\hat{\boldsymbol{\Theta}}^{(s+1)} := \arg\max_{\Theta} \mathcal{Q}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}^{(s)})$. The iteration process is repeated until convergence on the log-likelihood is met.

## 5.1 Complete log-likelihood

The complete log-likelihood can be written, up to some constant $c$, as:

$$\log \mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \ell_{ik} \Bigg[ \log(\pi_k) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\Phi_k|) - \frac{T}{2} \log(|\Sigma_k|) - \frac{1}{2} tr[\Sigma_k^{-1}(Z_i - M_k)\Phi_k^{-1}(Z_i - M_k)^\intercal] \Bigg] + c. \quad (3)$$

The unknown parameters to be estimated are $\boldsymbol{\Theta} := \{\pi_k, M_k, \Phi_k, \Sigma_k\}_{k=1}^K$

## 5.2 E-step

Looking at 3, keeping in time the block-structure of $Z_i$ and the links we defined in 4, it is easy to see that the expected values to be computed are $\mathbb{E}(\ell_{ik} | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$, $\mathbb{E}(\ell_{ik} Z_i^\beta | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ and of $\mathbb{E}(\ell_{ik} Z_i^\beta \Phi_k^{-1(s)} Z_i^{\beta\intercal} | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ or $\mathbb{E}(\ell_{ik} Z_i^{\beta\intercal} \Sigma_k^{-1(s)} Z_i^\beta | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ by the cyclic property of the trace. We will compute both as they are both needed in the M-step.

The first involves computing a cumulative probability of a matrix-variate normal distribution according to the thresholds described in Section 4. This in turn means solving a complex high-dimensional integral, which is hardly tractable analytically. However, it can be approximated through a Monte-Carlo approach applied on the vectorized reparametrization of the matrix-variate distribution according to Section 3.

The remaining three require the computation of the first and second moments of a truncated matrix-variate distribution. However, again that is a complex task with no close

solution, so we will need to work the issue around. We can bypass the problem by again working on the vectorized version of the distribution through the use of a Monte Carlo approach and specifically the use of a Gibbs sampler to sample from a truncated multivariate normal distribution. The samples generated to calculate the first moment can be reused to compute the second moment by calculating the inner product of the vectors used to compute the first then calculating the sample mean of these inner products.

### 5.3 M-step

To maximize the expected complete log-likelihood we can take the derivatives of Equation 3 with respect to the parameters. All updating equations have closed form and can be computed thanks to the expectations found in the E-step.

## 6 Conclusions

Mixture of matrix-variate normal distributions can be an efficient way to cluster longitudinal continuous data. Assuming that non-continuous variables are a discretization of latent continuous variables allows us to extend the use of these MMN to cluster longitudinal mixed data sets. Numerical study on synthetic data sets as well as real data application concerning diet choice during the pandemic (François-Lecompte et al., 2020) will be presented.

## References


[1] Francesco Amato, Julien Jacques, and Isabelle Prim-Allaz. "Clustering longitudinal ordinal data via finite mixture of matrix-variate distributions". In: *Statistics and Computing* 34.2 (Apr. 2024). ISSN: 1573-1375. DOI: 10.1007/s11222-024-10390-z.

[2] Kaye E. Basford and Geoffrey J. McLachlan. "The mixture method of clustering applied to three-way data". In: *Journal of Classification* 2.1 (Dec. 1985), pp. 109–125. ISSN: 1432-1343. DOI: 10.1007/BF01908066.

[3] Charles Bouveyron et al. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge, England, UK: Cambridge University Press, June 2019. ISBN: 978-1-10864418-1. DOI: 10.1017/9781108644181.

[4] Young-Geun Choi, Soohyun Ahn, and Jayoun Kim. "Model-Based Clustering of Mixed Data With Sparse Dependence". In: *IEEE Access* 11 (July 2023), pp. 75945–75954. DOI: 10.1109/ACCESS.2023.3296790.

[5] Marco Corneli, Charles Bouveyron, and Pierre Latouche. "Co-Clustering of Ordinal Data via Latent Continuous Random Variables and Not Missing at Random Entries". In: *Journal of Computational and Graphical Statistics* 29.4 (Oct. 2020), pp. 771–785. ISSN: 1061-8600. DOI: 10.1080/10618600.2020.1739533.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (Sept. 1977), pp. 1–22. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[7] Agnès François-Lecompte et al. "Confinement et comportements alimentaires - Quelles évolutions en matière d'alimentation durable ?" In: *Revue Française de Gestion* 46.293 (Nov. 2020), pp. 55–80. ISSN: 0338-4551. DOI: 10.3166/rfg.2020.00493.

[8] Michael P. B. Gallaugher and Paul D. McNicholas. "Finite mixtures of skewed matrix variate distributions". In: *Pattern Recognition* 80 (Aug. 2018), pp. 83–93. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2018.02.025.

[9] Arjun Kumar Gupta and Daya Krishna Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 2000.

[10] Damien McParland and Isobel Claire Gormley. "Model based clustering for mixed data: clustMD". In: *Advances in Data Analysis and Classification* 10.2 (June 2016), pp. 155–169. ISSN: 1862-5355. DOI: 10.1007/s11634-016-0238-x.

[11] Volodymyr Melnykov and Xuwen Zhu. "On model-based clustering of skewed matrix data". In: *Journal of Multivariate Analysis* 167 (Sept. 2018), pp. 181–194. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2018.04.007.

[12] Volodymyr Melnykov and Xuwen Zhu. "Studying crime trends in the USA over the years 2000–2012". In: *Advances in Data Analysis and Classification* 13.1 (Mar. 2019), pp. 325–341. ISSN: 1862-5355. DOI: 10.1007/s11634-018-0326-1.

[13] Shuchismita Sarkar et al. "On parsimonious models for modeling matrix data". In: *Computational Statistics & Data Analysis* 142 (Feb. 2020), p. 106822. ISSN: 0167-9473. DOI: 10.1016/j.csda.2019.106822.

[14] Margot Selosse et al. "Analysing a quality-of-life survey by using a co-clustering model for ordinal data and some dynamic implications". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.5 (Nov. 2019), pp. 1327–1349. ISSN: 0035-9254. DOI: 10.1111/rssc.12365.

[15] Cinzia Viroli. "Finite mixtures of matrix normal distributions for classifying three-way data". In: *Statistics and Computing* 21.4 (Oct. 2011), pp. 511–522. ISSN: 1573-1375. DOI: 10.1007/s11222-010-9188-x.

[16] Cinzia Viroli. "Model based clustering for three-way data structures". In: *Bayesian Analysis* 6.4 (Dec. 2011), pp. 573–602. ISSN: 1936-0975. DOI: 10.1214/11-BA622.

[17] Yang Wang and Volodymyr Melnykov. “On variable selection in matrix mixture modelling”. In: *Stat* 9.1 (Jan. 2020), e278. ISSN: 2049-1573. DOI: 10.1002/sta4.278.

[18] Xuwen Zhu, Shuchismita Sarkar, and Volodymyr Melnykov. “MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling”. In: *Journal of Classification* 39.1 (Mar. 2022), pp. 147–170. ISSN: 1432-1343. DOI: 10.1007/s00357-021-09401-9.

# Prédiction dynamique non paramétrique d'un risque d'événement à partir de prédicteurs longitudinaux

Corentin Ségalas [1,*], Cécile Proust-Lima [2] & Robin Genuer [1]

[1] *Univ. Bordeaux, INSERM, INRIA, BPH, U1219, F-33000 Bordeaux, France*
[2] *Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France*
[*] *corentin.segalas@u-bordeaux.fr*

**Résumé.** Prédire dynamiquement un risque de survenue d'évènement en prenant en compte l'historique médical complet d'un patient représente un défi statistique. En effet, de tels prédicteurs incluent souvent des variables qui évoluent au cours du temps et pour lesquelles on ne possède que des observations bruitées, mesurées à des temps irréguliers. Les approches proposées dans la littérature ont d'importantes limites. L'estimation des modèles conjoints devient impossible lorsque le nombre de prédicteurs longitudinaux croît et l'approche par regression calibration en deux étapes ignore la présence de données manquantes informatives. On propose une approche totalement non paramétrique, robuste aux données manquantes et qui permet d'inclure un grand nombre de prédicteurs longitudinaux, potentiellement mesurés irrégulièrement. Cette nouvelle méthode combine le principe des forêts aléatoires de survie (capables de gérer naturellement l'aspect grande dimension et la prédiction dynamique) avec l'analyse en composantes principales fonctionnelles (qui permet de résumer la dynamique temporelle des marqueurs).

**Mots-clés.** Analyse en composantes principales fonctionnelles, Données longitudinales, Forêts aléatoires, Prédiction dynamique.

**Abstract.** Dynamic prediction of the risk of an event taking into account a patient's complete medical history represents a statistical challenge. Indeed, such predictors often include variables which evolve over time and for which we only have noisy observations, measured at irregular times. The approaches proposed in the literature have important limitations. The estimation of joint models becomes impossible when the number of longitudinal predictors increases and the two-step regression calibration approach ignores the presence of informative missing data. We propose a completely non-parametric approach, robust to missing data and which allows the inclusion of a large number of longitudinal predictors, measured irregularly. This new method combines the principle of survival random forests (capable of naturally managing the high dimension aspect and dynamic prediction) with functional principal component analysis (which allows the temporal dynamics of the markers to be taken into account).

**Keywords.** Functional principal component analysis, Longitudinal data, Non parametric dynamic prediction, Random forests.

# 1 Introduction

En santé, il est courant de vouloir prédire le risque individuel de survenue d'un événement à partir de l'historique médical d'un patient. Implémenter de tels modèles représente un challenge statistique car ils doivent pouvoir incorporer un grand nombre de prédicteurs, dont certains évoluent au cours du temps. De plus, dans les études de santé, ces prédicteurs sont souvent observés de manière irrégulière, avec erreur de mesure et leur trajectoire temporelle peut être tronquée par la survenue de l'évènement. Les méthodes classiques, basées sur le modèle de Cox par exemple, ne sont adaptées à la prise en compte de tels prédicteurs.

Dans la littérature, trois grandes approches ont été proposées pour prédire un risque d'événement à partir de données longitudinales :

- l'approche par regression calibration [1] qui modélise séparément les trajectoires des prédicteurs puis les inclut dans des modèles de prédiction. Le problème de cette approche est que ce fonctionnement en deux étapes présente un biais en cas de données manquantes informatives.
- l'approche landmark [10] qui se place à un temps de prédiction et utilise tout l'historique jusqu'à ce temps pour prédire la survenue ultérieure de l'évènement. Ici, la censure informative est prise en compte mais au prix d'une réduction de l'information disponible.
- l'approche par modèle conjoint [8] qui modélise simultanément les trajectoires des prédicteurs et le risque d'évènement. Mais l'estimation des modèles conjoints devient trop lourde quand le nombre de prédicteurs longitudinaux croît.

Les forêts aléatoires [2] constituent un modèle prédictif dont l'avantage est de pouvoir modéliser des relations complexes et non linéaires entre prédicteurs en agrégeant un ensemble d'arbres de décision. Un arbre effectue un partitionnement récursif binaire de l'espace des prédicteurs en régions de plus en plus homogènes (en terme de variable à prédire). Elles ont été étendues au contexte de l'analyse de survie [5] mais sans pouvoir inclure de prédicteurs longitudinaux. Dans un précédent travail [3], les forêts dynamiques de survie ont été développées. Au sein de chaque noeud et pour chaque arbre de la forêt, chaque prédicteur longitudinal sélectionné comme candidat potentiel pour le calcul de la séparation optimale est modélisé par un modèle mixte [6], puis résumé à l'aide des effets aléatoires individuels prédits. De fait, cette approche est paramétrique et nécessite de spécifier les modèles mixtes. Cela peut impacter le temps de calcul du modèle et ne permet pas d'envisager des prédicteurs longitudinaux mesurés de façon intensive. Dans ce travail, on étend la méthode des forêts dynamiques en résumant les prédicteurs longitudinaux par les scores individuels issus d'une analyse en composante principale fonctionnelle [7] conduisant à une approche complètement non-paramétrique.

# 2 Méthode

Pour chaque participant $i \in \{1, \ldots, N\}$, on note $T_i$ le temps de survenue de l'évènement, $C_i$ le temps de censure supposé indépendant de $T_i$ et l'on observe alors $\tilde{T}_i = \min(T_i, C_i)$. On note $\delta_i$ l'indicateur d'évènement qui vaut $k \in \{1, \ldots, K\}$ si l'évènement de cause $k$ est

observé, 0 s'il est censuré. On collecte également $P$ prédicteurs indépendants du temps, $X_{ip}$ avec $p \in \{1, \ldots, P\}$ et $Q$ prédicteurs dépendants du temps $Y_{ijq}$ avec $j \in \{1, \ldots, n_{iq}\}$ et $q \in \{1, \ldots, Q\}$ mesurés aux temps $t_{ijm} \leq \tilde{T}_i$.

Le principe de la forêt aléatoire est le suivant : la prédiction finale s'obtient en agrégeant les prédictions d'un ensemble d'arbres de décision. Chaque arbre est construit sur un échantillon bootstrap de l'échantillon initial et procède par partitionnement binaire récursif de l'espace des prédicteurs. La séparation choisie est celle qui maximise la distance (en terme de variable à prédire) entre les deux sous-groupes. Les forêts aléatoires incluent un aléa supplémentaire en ne considérant comme candidats à la division qu'une sélection aléatoire de ces prédicteurs. La méthode des forêts aléatoires a été étendue à l'analyse de survie en utilisant comme distance la statistique du logrank [5] et en proposant une méthode d'agrégation des arbres adaptée. La méthode est aussi applicable dans un contexte de risques compétitifs en considérant le test de Fine and Gray [4].

Néanmoins, ces dernières ne permettent pas d'inclure des prédicteurs dépendant du temps. C'est pourquoi les forêts dynamiques [3] ont été introduites. Le principe est le suivant : pour chacun des arbres de la forêt, au sein de chaque noeud et avant toute division, une fois la sélection aléatoires des prédicteurs effectuée, les trajectoires individuelles des prédicteurs longitudinaux $(Y_{ijq})$ vont être résumées par des quantités rendues indépendantes du temps : les effets aléatoires prédits après estimation d'un modèles mixte sur $(Y_{ijq})_{ij}$.

Dans ce travail, on conserve la même architecture de construction de la forêt aléatoire, mais en utilisant les scores issus de l'analyse en composante principale fonctionnelle (ACPF) pour résumer les trajectoires longitudinales plutôt que les effets aléatoires prédits par les modèles mixtes. L'avantage de cette approche est qu'elle évite de faire des hypothèses paramétriques sur les formes des trajectoires ou sur la distribution des effets aléatoires contrairement à ce qui a été proposé [3]. Si l'ACPF a été initialement développée pour des données fonctionnelles denses et régulières [7], l'algorithme PACE [11] a été introduit pour adapter l'ACPF à des données fonctionnelles éparses et irrégulières.

Plaçons nous dans le cadre de l'analyse de données fonctionnelles et supposons que, pour un $q \in \{1, \ldots, Q\}$ fixé, les trajectoires $(Y_{ijq})_{ij}$ sont la collection bruitée de réalisations aléatoires $(f_{iq})_i$ d'une fonction inconnue sous-jacente $f_q$. L'ACPF est basée sur la décomposition de Karhunen-Loève qui, sous des hypothèses de régularité, assure que

$$f_{iq}(t) = \mu_q(t) + \sum_{k=1}^{\infty} \xi_{iqk}\phi_{qk}(t)$$

où $\xi_{ikq}$ et $\phi_{qk}(t)$ sont respectivement les valeurs propres (aussi appelés scores de l'ACPF) et les fonctions propres issues de la décomposition en valeur propre de l'opérateur de covariance de $f_q$. Afin de réduire la dimension infinie de cette décomposition, on peut ne conserver que les $K$ premiers termes de la somme et les scores $\xi_{iqk}$ de l'ACPF représentent alors les coordonnées dans cet espace fonctionnel de dimension fini. Pour un individu $i$, ils mesurent la déviation individuelle à la fonction moyenne $\mu_q(t)$ et constituent un résumé intéressant de sa dynamique temporelle. $K$ peut être fixé arbitrairement ou en se basant sur des critères de pourcentage de variance expliquée. L'algorithme PACE [11] permet directement d'obtenir des estimations de ces scores $\hat{\xi}_{iqk}$ mais aussi de la fonction moyenne $\hat{\mu}_q(\mathbf{t})$ et des composantes

principales fonctionnelles $\hat{\phi}_{qk}(\mathbf{t})$ sur une grille de temps $\mathbf{t}$.

Ainsi, au sein de chaque noeud et avant toute division, une fois les prédicteurs candidats sélectionnés (parmi les $P$ indépendants du temps et les $Q$ dépendants du temps), on applique aux candidats longitudinaux l'algorithme PACE. Pour chaque participant $i \in \{1, \ldots, N\}$, on ne disposera alors que de prédicteurs indépendants du temps (des variables indépendantes du temps et des scores estimés par l'ACPF $\hat{\xi}_{iqk}$). On peut maintenant appliquer la stratégie usuelle des forêts aléatoires de survie [5] puis effectuer la division optimale avec pour critère la statistique du log-rank. La construction de l'arbre se termine lorsque le critère d'arrêt est atteint (le nombre d'évènements minimal dans un noeud est atteint par exemple). À ce stade, on suppose que les noeuds terminaux sont suffisamment homogènes en terme de probabilité de survenue de l'évènement. Dans chacun de ces noeuds terminaux, la fonction de risque cumulé est estimée par l'estimateur de Nelson-Aalen.

Pour un nouvel individu, il est possible de faire de la prédiction dynamique de survenue de l'évènement à partir d'un temps de prédiction $s$ pour un horizon donné. Pour cela, il faut tout d'abord tronquer pour cet individu les prédicteurs longitudinaux au temps $s$. Puis on peut prédire avec la forêt l'évènement à partir de ces données. Cela nécessite, au sein de chaque noeud de calculer les coordonnées de cet individu dans la base de l'espace fonctionnel de dimension $K$ obtenue sur l'échantillon d'apprentissage. L'estimation finale s'obtient alors en agrégeant les probabilités prédites de l'évènement dans les feuilles où le sujet est classé.

# 3 Simulation et application

À la différence des modèles mixtes, connus pour être robustes aux données manquantes dites *missing at random* (la probabilité qu'un marqueur ne soit pas observé ne dépend pas des valeurs non observées de ce marqueur), le comportement de l'ACPF en présence de données manquantes est incertain. Nous avons mené une étude de simulation préliminaire [9] qui a permis d'assurer que l'ACPF, sous divers scénarios de données manquantes, se comportait aussi bien qu'un modèle mixte à l'exception des scénarios les plus extrêmes. Cet argument, associé au fait que, par l'architecture de la forêt aléatoire de survie, les profils de survie des individus tendent à se ressembler, vont dans le sens d'une bonne robustesse aux données manquantes *missing at random*.

Nous avons réalisé une étude de simulation pour valider cette nouvelle approche en étudiant l'impact des données manquantes et de la stratégie de choix de $K$. On s'est aussi intéressé à la prise en compte de la variabilité de la trajectoire temporelle. Pour cela, deux stratégies ont été envisagées pour résumer l'information longitudinale : ACPF sur la trajectoire uniquement, ACPF sur la trajectoire et sa dérivée. Enfin, nous avons comparé cette approche à l'approche des forêts dynamiques par modèles mixtes proposée précédemment [3]. Nous présenterons les résultats de cette étude de simulation.

Nous présenterons également une application sur données réelles. L'objectif était de prédire le risque de survenue de vasospasme chez les patients hospitalisés après une hémorragie sub-arachnoïdienne (HSA). Il s'agit d'une complication majeure qui a lieu entre 3 et 14 jours après la survenue de la HSA, souvent détecté trop tard pour un traitement efficace. Dans les

facteurs de risque identifiés, certains sont monitorés au cours de l'hospitalisation de ces patients (par exemple la pression artérielle, l'hyperglycémie), toutes les heures pour la plupart. Ainsi l'approche des forêts dynamiques que nous avons développée est adaptée à ce genre de contexte et permettrait de construire un modèle prédictif utile pour aider le clinicien dans le suivi des patients.

# 4 Conclusion

Nous avons étendu la méthode des forêts dynamiques [3] et proposé une nouvelle approche totalement non paramétrique qui permet de faire de la prédiction dynamique du risque de survenue d'un évènement en incluant des prédicteurs, potentiellement en grand nombre, dépendant du temps, mesurés irrégulièrement et qui peuvent inclure des données manquantes. De plus, cette approche, basée sur la méthode des forêts aléatoires de survie, bénéficie d'outils statistiques qui permettent d'informer et d'évaluer l'importance des variables dans la prédiction. Cela permet d'éclairer les mécanismes de prédiction du modèle, évitant l'écueil de l'effet *boîte noire* de certains algorithmes prédictifs.

# Références


[1] Paul S. Albert et Joanna H. Shih : On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. *Biometrics*, 66(3):983–987, 2010.

[2] Leo Breiman : Random Forests. *Machine Learning*, 45(1):5–32, octobre 2001.

[3] Anthony Devaux, Catherine Helmer, Robin Genuer et Cécile Proust-Lima : Random survival forests with multivariate longitudinal endogenous covariates. *Statistical Methods in Medical Research*, octobre 2023.

[4] Hemant Ishwaran, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange et Bryan M. Lau : Random survival forests for competing risks. *Biostatistics (Oxford, England)*, 15(4):757–773, octobre 2014.

[5] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone et Michael S. Lauer : Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, septembre 2008.

[6] Geert Molenberghs et Geert Verbeke : *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer, New York, NY, 2000.

[7] James O. Ramsay et Bernard W. Silverman : *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, NY, 2005.

[8] Dimitris Rizopoulos : *Joint Models for Longitudinal and Time-to-Event Data : With Applications in R*. CRC Press, juin 2012.

[9] Corentin Ségalas, Catherine Helmer, Robin Genuer et Cécile Proust-Lima : Functional principal component analysis as an alternative to mixed-effect models for describing sparse repeated measures in presence of missing data, février 2024. arXiv :2402.10624 [stat].

[10] Hans C. Van Houwelingen : Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

[11] Fang Yao, Hans-Georg Müller et Jane-Ling Wang : Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

# Données directionnelles

# Inférence asymptotique pour des données directionnelles bruitées

Diego Bolon[1], Davy Paindaveine[2] & Thomas Verdebout[3]

[1] *Universidade de Santiago de Compostela, Espagne, diego.bolon.rodriguez@usc.es*
[2] *Université libre de Bruxelles, Belgique, Davy.Paindaveine@ulb.be*
[3] *Université libre de Bruxelles, Belgique, Thomas.Verdebout@ulb.be*

**Résumé.** Nous introduisons des modèles paramétriques pour des données directionnelles bruitées dans lesquels un bruit radial de magnitude $\sigma^2$ fait dévier les observations de leur domaine sphérique théorique, à savoir une hypersphère centrée en $\theta$ et de rayon $r$. Nous considérons les problèmes d'inférence — les tests d'hypothèses, l'estimation ponctuelle et l'estimation par zone de confiance — sur le paramètre de position $\theta$, dans un contexte où $r$ et $\sigma^2$ restent non spécifiés. Nous introduisons divers scénarios asymptotiques dans lesquels le rayon de l'hypersphère et, de façon plus importante, la magnitude du bruit peuvent dépendre de la taille d'échantillon $n$ d'une façon essentiellement arbitraire. Ceci nous permet de considérer des situations très diverses dans lesquelles l'information a priori que les données appartiennent à une hypersphère est de plus en plus, ou au contraire de moins en moins, pertinente. Nous basons notre étude sur la théorie asymptotique des expériences asymptotiques de Le Cam et notre objectif est d'obtenir une compréhension complète des expériences limites qui en résultent. Les taux de contiguïté associés, qui caractérisent la difficulté des problèmes d'inférence considérés, révèlent des résultats assez contre-intuitifs dans certains des scénarios traités. Nous construisons des estimateurs et des tests qui sont localement asymptotiquement optimaux de façon adaptative à travers les différents régimes. Nous montrons que, dans les scénarios asymptotiques standards, les procédures classiques qui ignorent l'information d'a priori hypersphérique réalisent le taux de convergence optimal mais n'atteignent pas les bornes d'efficacité, et que, dans des scénarios asymptotiques non-standards, ces procédures classiques n'ont pas le bon taux de convergence. Nous étudions par l'intermédiaires d'exercices de Monte Carlo à quel point nos résultats asymptotiques se matérialisent dans des échantillons de taille finie. Les perspectives de recherche future incluent notamment l'extension au cas non paramétrique dans lequel la loi du bruit n'est pas spécifiée.

**Mots-clés.** Théorie asymptotique des expériences statistiques de Le Cam, double asymptotique, taux de contiguïté, données presque directionnelles, identifiabilité forte

**Abstract.** We introduce parametric models for noisy directional data, in which a radial noise with magnitude $\sigma^2$ makes the observations deviate from their theoretical hyperspherical sample space, namely a hypersphere centered at $\theta$ and with radius $r$. We consider inference — hypothesis testing, point estimation, and confidence zone estimation — on the location parameter $\theta$, in a framework where both $r$ and $\sigma^2$ remain unspecified. We introduce several asymptotic scenarios in which the radius of the hypersphere and, most importantly, the noise magnitude may depend on the sample size $n$ in an essentially arbitrary way. This allows

us to consider very diverse cases, in which the a priori information that the data belong to a hypersphere is more and more, or on the contrary less and less, relevant. We base our investigation on Le Cam's asymptotic theory of statistical experiments and aim at a full understanding of the resulting limiting experiments. The corresponding contiguity rates, that characterize how easy/hard inference on $\theta$ is, reveal rather counter-intuitive results in some scenarios. We build locally asymptotically optimal tests and estimators, that turn out to be adaptively optimal across all asymptotic scenarios. We show that, in standard asymptotic scenarios, classical procedures that would ignore the hyperspherical a priori information are rate-consistent but do not achieve efficiency bounds, and that, in non-standard asymptotic scenarios, such classical procedures are not even rate-consistent. We investigate the finite-sample relevance of our results through Monte Carlo exercises. Perspectives for future research in particular include the extension to the nonparametric case in which the distribution of the radial noise remains unspecified.

**Keywords.** Le Cam's theory of asymptotic experiments, contiguity rates, double asymptotics, nearly directional data, strong identifiability

# 1 Contexte

Les problèmes de position multivariés comptent probablement parmi les problèmes les plus étudiés en statistique. Les tests les plus standards pour les problèmes de position à un et deux échantillons sont certainement les tests de Hotelling; voir Hotelling (1931). Ils sont encore beaucoup étudiés aujourd'hui; nous renvoyons par exemple à Chen et al. (2011) et Feng et al. (2017) pour des tests de Hotelling régularisés à un échantillon, et à Li et al. (2020) pour des tests d'une nature similaire pour le problème à deux échantillons. Les tests de Hotelling sont basés sur des moyennes arithmétiques, donc ne sont pas robustes à d'éventuelles observations aberrantes ou à des queues lourdes. Ceci a motivé l'introduction de tests non paramétriques, et en particulier de tests de signes et de rangs signés; voir, parmi beaucoup d'autres, Randles (2000), Hallin et Paindaveine (2002), Larocque, Nevalainen et Oja (2007), Wang, Peng et Li (2015), et Feng, Zou et Wang (2016). Des contributions récentes en inférence pour la position multivariée incluent Agostinelli et Greco (2019) qui a étudié l'estimation par vraisemblance pondérée, Frahm, Nordhausen et Oja (2020), qui a proposé des estimateurs de position pour des données incomplètes et dépendantes, Dürre et Paindaveine (2022), qui a défini des estimateurs de position affine-équivariants fondés sur des simplexes, Chakraborty et Chaudhuri (2017), Kock et Preinerstorfer (2019, 2023), qui ont considéré le cas des grandes dimensions, et Ley et al. (2013), Paindaveine et Verdebout (2017, 2020a,b), qui ont traité les problèmes de position dans le contexte des données directionnelles.

Que ce soit en faible ou en grande dimension, il est de plus en plus courant de supposer que la dimension effective des données, $k$ disons, est plus petite que la dimension $d$ de l'espace ambient; pour ce qui est la grande dimension, nous renvoyons par exemple à Wright et Ma (2022). Sous une hypothèse de linéarité qui veut que les données proviennent d'une version bruitée d'une distribution concentrée sur un hyperplan de dimension $k$ de $\mathbb{R}^d$, la méthodologie

classique dans ce cadre repose alors sur l'analyse en composantes principales, mais bien sûr il est plus prometteur et général de considérer des techniques de réduction de la dimension non linéaires qui permettent aux données de dévier d'une variété $k$-dimensionnelle courbée. En pratique, tant la dimension $k$ intrinsèque que la variété correspondante restent non spécifiées, et une vaste littérature a considéré le problème de reconstruire ces quantités-clés inconnues; nous renvoyons, par exemple, à Donoho et Grimes (2003), Maggioni, Minsker et Strawn (2016), et aux travaux cités dans ces articles. Des articles récents étudiant l'inférence statistique pour des données déviant de façon modérée d'une variété incluent notamment Shapiro, Xie et Zhang (2021) et Cheng et Xie (2024), qui ont respectivement considéré des tests de goodness-of-fit et des tests à deux échantillons.

Typiquement, les résultats obtenus dans la littérature fournissent des vitesses de convergence et, au mieux, des résultats d'optimalité de type minimax sous des conditions convenables; voir, par exemple, Shapiro, Xie et Zhang (2021) et Cheng et Xie (2024). Pour autant que nous sachions, des résultats d'optimalité/efficacité plus fins, qui établiraient par exemple une optimalité au sens de Le Cam, n'ont pas été obtenus dans ce cadre. Notre travail a pour objectif d'obtenir de tels résultats d'optimalité. Puisqu'il n'y a pas de "free lunch", le prix à payer est de faire des hypothèses plus fortes sur la forme de la variété sous-jacente, par exemple, de supposer que cette variété est une hypersphère ou un hypertore. Dans cet exposé, nous suivons en effet cette route et adoptons un cadre dans lequel l'échantillon à disposition est fait de versions bruitées de vecteurs aléatoires prenant leurs valeurs sur une hypersphère de $\mathbb{R}^d$.

# Bibliographie


Agostinelli, C. et Greco, L. (2019), Weighted likelihood estimation of multivariate location et scatter, *Test*, 28, pp. 756-784.

Chakraborty, A. et Chaudhuri, P. (2017), Tests for high-dimensional data based on means, spatial signs and spatial ranks, *Annals of Statistics*, 45, pp. 771-799.

Chen, L. S., Paul, D., Prentice, R. L. et Wang, P. (2011), A regularized Hotelling's $T^2$ test for pathway analysis in proteomic studies, *Journal of American Statistical Association*, 106, pp. 1345-1360.

Cheng, X., et Xie, Y. (2024), Kernel two-sample tests for manifold data, *Bernoulli*. A paraître.

Kock, A. B. et Preinerstorfer, D. (2019), Power in high-dimensional testing problems, *Econometrica*, 87, pp. 1055-1069.

Donoho, D., et Grimes, C. (2003), Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Science USA*, 100, pp. 5591-5596.

Dürre, A. et Paindaveine, D. (2022) Affine-equivariant inference for multivariate location under $L_p$ loss functions, *Annals of Statistics*, 50, pp. 2616-2640.

Feng, L., Zou, C., et Wang, Z. (2016), Multivariate-sign-based high-dimensional tests for the two-sample location problem, *Journal of American Statistical Association*, 111, pp. 721-735.

Feng, L., Zou, C., Wang, Z., et Zhu, L. (2017), Composite $T^2$ test for high-dimensional data, *Statistica Sinica*, 27, pp. 1419-1436.

Frahm, G., Nordhausen, K. et Oja, H. (2020), M-estimation with incomplete and dependent multivariate data, *Journal of Multivariate Analysis*, 176, pp. 104569.

Hallin, M. et Paindaveine, D. (2002), Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks, *Annals of Statistics*, 30, pp. 1103-1133.

Hotelling, H. (1931), The generalization of Student's ratio, *Annals of Mathematical Statistics*, 2, pp. 360-378.

Kock, A. B., et Preinerstorfer, D. (2023), Consistency of $p$-norm based tests in high dimensions: Characterization, monotonicity, domination, *Bernoulli*, 29, pp. 2544–2573.

Larocque, D., Nevalainen, J. et Oja, H. (2007), A weighted multivariate sign test for cluster-correlated data *Biometrika*, 94, pp. 267-283.

Ley, C., Swan, Y., Thiam, B. et Verdebout, T. (2013), Optimal R-estimation of a spherical location *Statistica Sinica*, 23, pp. 305–333.

Li, H., Aue, A. Paul, D., Peng, J. et Wang, P. (2020), An adaptable generalization of Hotelling's $T^2$ test in high dimension *Annals of Statistics*, 48, pp. 1815-1847.

Maggioni, M., Minsker, S. et Strawn, N. (2016) Multiscale dictionary learning: non-asymptotic bounds and robustness, *Journal of Machine Learning Research*, 17, pp. 43-93.

Paindaveine, D. et Verdebout, T. (2017), Inference on the mode of weak directional signals: a Le Cam perspective on hypothesis testing near singularities, *Annals of Statistics*, 45, pp. 800-832.

Paindaveine, D. et Verdebout, T. (2020a), Detecting the direction of a signal on high-dimensional spheres: non-null and Le Cam optimality results *Probability Theory and Related Fields*, 176, pp. 1165-1216.

Paindaveine, D. et Verdebout, T. (2020b), Inference for spherical location under high concentration, *Annals of Statistics*, 48, pp. 2982-2998.

Randles, R.H. (2000), A simpler, affine-invariant, multivariate, distribution-free sign test, *Journal of American Statistical Association*, 95, pp. 1263-1268.

Shapiro, A., Xie, Y. et Zhang, R. (2021), Goodness-of-fit tests on manifolds, *IEEE Transactions on Information Theory*, 67, pp. 2539-2553.

Wang, L., Peng, B. et Li, R. (2015), A high-dimensional nonparametric multivariate test for mean vector, *Journal of American Statistical Association*, 110, pp. 1658-1669.

Wright, J. et Ma, Y. (2022), *High-Dimensional Data Analysis with Low-Dimensional Models. Principles, Computation, and Applications*, Cambridge University Press.

# Comportement asymptotique de tests de Sobolev sur la sphère unité.

Eduardo García-Portugués, Davy Paindaveine and Thomas Verdebout

*ECARES and Mathematics Department, Université Libre de Bruxelles, Boulevard du Triomphe, CP210, B-1050 Brussels, Belgium,* email: tverdebo@ulb.ac.be.

**Résumé.** L'un des problèmes les plus classiques de la statistique multivariée est le problème de test d'uniformité sur l'hypersphère unité. Plutôt que de se limiter à des tests qui ne peuvent détecter que des types spécifiques d'alternatives, nous considérons la vaste classe des tests de Sobolev. Bien que certains de ces tests soient connus pour être "omnibus", leur comportement asymptotique sous l'alternative ainsi que leurs taux de détection, de manière inattendue, restent largement inexplorés. Pour améliorer cette situation, nous étudions en détail les puissances asymptotiques locales des tests de Sobolev dans le cadre d'alternatives classiques à l'uniformité, à savoir les alternatives à symétrie rotationelle. Nous montrons en particulier que les taux de détection des tests de Sobolev ne dépendent pas seulement des coefficients qui définissent ces tests, mais aussi des dérivées de la fonction angulaire sous-jacente en zéro.

**Mots-clés.** Données directionnelles, tests d'uniformité, tests de Sobolev.

**Abstract.** One of the most classical problems in multivariate statistics is considered, namely, the problem of testing isotropy, or equivalently, the problem of testing uniformity on the unit hypersphere. Rather than restricting to tests that can detect specific types of alternatives only, we consider the broad class of Sobolev tests. While these tests are known to allow for omnibus testing of uniformity, their non-null behavior and consistency rates, unexpectedly, remain largely unexplored. To improve on this, we thoroughly study the local asymptotic powers of Sobolev tests under the most classical alternatives to uniformity, namely, under rotationally symmetric alternatives. We show in particular that the consistency rate of Sobolev tests does not only depend on the coefficients defining these tests but also on the derivatives of the underlying angular function at zero.

**Keywords.** Directional data, uniformity tests, Sobolev tests.

## 1 Directional data and testing for uniformity

Directional statistics are dealing with observations that belong to the unit hypersphere $\mathbb{S}^{p-1} := \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|^2 = \mathbf{u}'\mathbf{u} = 1\}$ of $\mathbb{R}^p$ or more generally on compact Riemannian manifolds. Instances of directional data happen in meteorology (wind directions), astronomy (directions of cosmic rays, positions of stars), paleomagnetism (remanence directions), biology (protein structure, studies of animal navigation), forest sciences (directions of wildfire

propagation), medicine (head normal vectors), and text mining (quantitative representation of documents in high-dimensional hyperspheres), to cite but some. Classical monographs on directional statistics are Watson (1983) and Mardia and Jupp (2000); a recent book that overviews the usage of some modern methods in directional statistics is Ley and Verdebout (2017).

When modeling directional data, that is, unit-norm multivariate vectors, a first natural question is to ask whether the directions at hand are uniformly distributed or, on the contrary, whether there exist modes of variation significantly different from uniformity. On the basis of $n$ i.i.d. observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$ with common distribution P on $\mathbb{S}^{p-1}$, the problem we tackle in this work is the problem of testing $\mathcal{H}_0 : \mathrm{P} \equiv \mathrm{Unif}(\mathbb{S}^{p-1})$ against $\mathcal{H}_1 : \mathrm{P} \neq \mathrm{Unif}(\mathbb{S}^{p-1})$, where $\mathrm{Unif}(\mathbb{S}^{p-1})$ stand for the uniform probability measure on $\mathbb{S}^{p-1}$. We study in this work tests belonging to the class of Sobolev tests for this problem. Sobolev tests are introduced in the next Section.

## 2 Sobolev tests

The class of so-called *Sobolev tests* has been introduced by Beran (1968, 1969) and Gine (1975). Sobolev tests are obtained using the eigenfunctions of the *Laplace–Beltrami operator* (or *Laplacian*) $\Delta$ acting on $\mathbb{S}^{p-1}$. Using the $n$-tuple of observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$, a Sobolev test rejects the null hypothesis of uniformity $\mathcal{H}_0$ for large values of

$$S_n := \frac{1}{n} \sum_{i,j=1}^{n} \sum_{k=1}^{\infty} v_k^2 \langle \mathrm{t}_k(\mathbf{U}_i), \mathrm{t}_k(\mathbf{U}_j) \rangle, \tag{2.1}$$

where $\mathbf{u} \to \mathrm{t}_k(\mathbf{u})$ is a mapping from $\mathbb{S}^{p-1}$ to the space of eigenfunctions associated with the $k$th non-zero eigenvalue of the Laplacian, the $v_k$'s are weights and $\langle f, g \rangle := \int_{\mathbb{S}^{p-1}} f(\mathbf{u}) g(\mathbf{u}) \, \mathrm{d}\mu(\mathbf{u})$ denotes the inner product on $L^2(\mathbb{S}^{p-1}, \mu)$ ($\mu$ is the surface area measure on $\mathcal{S}^{p-1}$). An explicit form for $\langle \mathrm{t}_k(\mathbf{U}_i), \mathrm{t}_k(\mathbf{U}_j) \rangle$ on $\mathbb{S}^{p-1}$ exists. More precisely, given $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{p-1}$,

$$\langle \mathrm{t}_k(\mathbf{u}), \mathrm{t}_k(\mathbf{v}) \rangle = \begin{cases} 2\cos(k \angle(\mathbf{u}, \mathbf{v})), & \text{if } p = 2, \\ \left(1 + \frac{2k}{p-2}\right) C_k^{(p-2)/2}(\mathbf{u}'\mathbf{v}), & \text{if } p > 2, \end{cases} \tag{2.2}$$

where $\cos \angle(\mathbf{u}, \mathbf{v}) = \mathbf{u}'\mathbf{v}$ and $C_k^{\alpha}$ denote the Gegenbauer polynomial of index $\alpha$ and order $k$. Well-known Sobolev tests are

- the *Rayleigh test*. Taking $v_1 = 1$ and $v_k = 0$ for $k \geq 2$ in (2.1) we obtain the Rayleigh test statistic on $\mathbb{S}^{p-1}$ given by

$$R_n = \frac{p}{n} \sum_{i,j=1}^{n} \mathbf{U}_i' \mathbf{U}_j. \tag{2.3}$$

  Under $\mathcal{H}_0$, $R_n$ is asymptotically $\chi_p^2$ distributed.

- the *Bingham test*. When $\mathbf{U} \sim \text{Unif}(\mathbb{S}^{p-1})$, then $\mathbb{E}[\mathbf{U}\mathbf{U}'] = \frac{1}{p}\mathbf{I}_p$. The Bingham test evaluates this latter sphericity property of $\mathbf{U}$ by the test statistic

$$B_n := \frac{np(p+2)}{2}\left(\text{tr}(\mathbf{S}^2) - \frac{1}{p}\right),$$

where $\mathbf{S} := \frac{1}{n}\sum_{i=1}^n \mathbf{U}_i\mathbf{U}_i'$ is the empirical covariance matrix of the $\mathbf{U}_i$'s. Under $\mathcal{H}_0$, $B_n$ is asymptotically $\chi^2_{(p-1)(p+2)/2}$ distributed. The statistic $B_n$ is obtained by letting $v_2 = 1$ and $v_k = 0$ for $k \neq 2$ in (2.1).

While much is known about the asymptotic behavior of several Sobolev tests under the null hypothesis of uniformity and when the dimension is fixed, less is known about the asymptotic behaviour of such tests under local alternatives, even under the rotationally symmetric alternatives defined in the next section.

# 3 Asymptotic powers of Sobolev tests

We consider in this work specific alternatives to the null of uniformity over the $p$-dimensional unit sphere $\mathcal{S}^{p-1}$, namely rotationally symmetric alternatives. A $p$-dimensional unit vector $\mathbf{U}$ is called *rotationally symmetric about* $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ if and only if $\mathbf{O}\mathbf{U}$ is equal in distribution to $\mathbf{U}$ for any orthogonal $p\times p$ matrix $\mathbf{O}$ satisfying $\mathbf{O}\boldsymbol{\theta} = \boldsymbol{\theta}$. We actually restrict to rotationally symmetric densities of the form

$$\mathbf{u} \mapsto c_{p,\kappa,f} f(\kappa\, \mathbf{u}'\boldsymbol{\theta}), \qquad \mathbf{x} \in \mathcal{S}^{p-1}, \tag{3.4}$$

where $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ is a location parameter, $\kappa(>0)$ is a concentration parameter, and the function $f : \mathbb{R} \to \mathbb{R}^+$ is monotone strictly increasing, twice differentiable at 0, and satisfies $f(0) = f'(0) = 1$. Consider triangular arrays of observations $\mathbf{U}_{ni}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$ where the random vectors $\mathbf{U}_{ni}$, $i = 1, \ldots, n$ take values in $\mathcal{S}^{p_n-1}$. More specifically, for any $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$, $\kappa_n > 0$ and $f$ as above, we will denote as $\text{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}$ the hypothesis under which $\mathbf{U}_{ni}$, $i = 1, \ldots, n$ are mutually independent and share the common density $\mathbf{u} \mapsto c_{p_n,\kappa_n,f} f(\kappa_n\, \mathbf{u}'\boldsymbol{\theta}_n)$; $\text{P}^{(n)}_0$ will denote triangular arrays of uniformly distributed observations. We have the following result (note that in this result $p_n$ may diverges to $\infty$).

**Proposition 3.1** *Let $(p_n)$ be a sequence in $\{2, 3, \ldots\}$. Let $(\boldsymbol{\theta}_n)$ be a sequence such that $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$ for all $n$, $(\kappa_n)$ be a positive sequence such that $\kappa_n = O(\sqrt{\frac{p_n}{n}})$. Then, the sequence of alternative hypotheses $\text{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}$ and the null sequence $\text{P}^{(n)}_0$ are mutually contiguous.*

Proposition 3.1 provides the contiguous alternatives to the uniform. Some Sobolev tests described in the previous are known to allow for omnibus testing of uniformity. Their non-null behavior and consistency rates, unexpectedly, remain largely unexplored. In particular, nothing is know about their potential rate-consistency. A natural question is then "do Sobolev tests detect the contiguous alternatives of Proposition 3.1 above?" To tackle this question, we thoroughly study the local asymptotic powers of Sobolev tests under rotationally symmetric alternatives. We show in particular that the consistency rate of a Sobolev test does not only depend on the coefficients defining the test but also on the derivatives of the underlying angular function at zero.

# Bibliographie

Beran, R. J. (1968). Testing for uniformity on a compact homogeneous space. *Journal of Applied Probability*, 5, pp. 177-195.

Beran, R. J. (1969). Asymptotic theory of a class of tests for uniformity of a circular distribution. *Annals of Mathematical Statistics*, 40, pp. 1196-1206.

Cutting, C., Paindaveine, D., and Verdebout, T. (2017). Testing uniformity on high-dimensional spheres against monotone rotationally symmetric alternatives. *Annals of Statistics*, 45(3), pp. 1024-1058.

Ley, C. and Verdebout, T. (2017). Modern Directional Statistics. Chapman and Hall/CRC.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Watson, G. S. (1983). *Statistics on Spheres*, volume 6 of University of Arkansas Lecture Notes in the Mathematical Sciences. John Wiley & Sons, New York.

# Test de Runs pour données directionnelles : propriétés locales et optimalités asymptotiques.

Maxime Boucher [1] & Christian Franck [2] & Yuichi Gotto [3] & Thomas Verdebout [4]

[1] *Université Libre de Bruxelles (ULB), Belgique, maxime.boucher@ulb.be*
[2] *CREST et Université de Lille, France, christian.francq@ensae.fr*
[3] *Université de Kyushu, Japon, goto.yuichi.436@m.kyushu-u.ac.jp*
[4] *Université Libre de Bruxelles (ULB), Belgique, thomas.verdebout@ulb.be*

**Résumé.** Dans le travail présenté ici, on s'intéresse au problème de détection des corrélations sérielles dans un contexte de données directionnelles. Motivé par une application sur des données réelles concernant la localisation de tâches solaires au cours des décennies, on définit un concept de runs proprement adapté au contexte directionnel. On montre alors que ce test, basé sur les runs directionnels, possède des propriétés locales et asymptotiques dans le cas d'alternatives locales avec des dépendances sérielles. A l'aide de simulations Monte-Carlo, on expose les propriétés, pour des tailles d'échantillons finies, de notre test et son utilité dans le cadre de l'étude des localisations des tâches solaires pendant les dix derniers cycles solaires.

**Mots-clés.** runs, données directionnelles, dépendance sérielle, optimalité locale et asymptotique, randomness.

**Abstract.** In the present work, we tackle the problem of detecting serial correlation in the context of directional data. Motivated by a real data example involving sunspots locations, we define a concept of runs properly adapted to the directional context. We then show that tests based on the latter runs enjoy some local and asymptotic property against local alternatives with serial dependence. We compute the finite sample performances of our tests using Monte Carlo simulations and show their usefulness on a real data illustration that involves the analysis of sunspots locations for various solar cycles.

**Keywords.** runs, directional data, randomness test, serial dependence, local and asymptotic optimality, randomness.

# 1 Test de runs directionnels

## 1.1 Contexte Directionnel

Dans le cas univarié, on définit un run comme une suite consécutive d'observations du même signe. Si l'on dispose d'un échantillon univarié $X_1, \ldots, X_n$ avec pour paramètre de localisation $\theta$, on définit le signe de $X_t$ par $U_t = sign(X_t - \theta)$. Le nombre de runs dans un échantillon peut-être alors calculé de la manière suivante :

$$\sum_{t=2}^{n} U_t(\theta) U_{t-1}(\theta) = N_n(\theta) - \mathbb{E}[N_n(\theta)].$$

où $N_n(\theta) := 1 + \sum_{t=2}^{n} \mathbb{I}[U_t(\theta) \neq U_{t-1}(\theta)]$ est le nombre de runs. On peut donc alors construire un test de runs pour tester la randomness de l'échantillon à l'aide de l'asymptotique gaussien classique.

Dans le cas multivarié, la notion de signe est adaptée par le signe multivarié comme étant, pour un échantillon de $p$-vecteurs $X_1, \ldots, X_n$, les quantités :

$$U_t = \frac{X_t - \theta}{\|X_t - \theta\|}$$

On définit alors la notion de run dans le cas multivarié par :

$$R_1^{(n)} := \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} U_t'(\theta) U_{t-1}(\theta)$$

où $\theta$ est à nouveau le vecteur de localisation. La quantité $R_1^{(n)}$ permet alors de construire un test afin de détecter une possible dépendance dans l'échantillon (une dépendance avec un lag de taille 1 puisque l'on regarde l'angle entre deux vecteurs consécutifs).

Dans ce travail, on se focalise plus particulièrement sur les runs pour un échantillon directionnel. C'est-à-dire que l'on dispose d'un échantillon de $p$-vecteurs sur la sphère $\mathcal{S}^{p-1}$ de $\mathbb{R}^p$. Dans le cadre directionnel, les données peuvent se décomposer de la manière suivante:

$$X_t = v_t(\theta)\theta + \sqrt{1 - v_t(\theta)^2}\Gamma_\theta S_t(\theta)$$

où $v_t(\theta)$ est la partie projetée de $X_t$ sur la direction $\theta$, $\Gamma_\theta$ est une matrice semi-orthogonale qui vérifie $\Gamma_\theta \Gamma_\theta' = I_p - \theta\theta'$ et $\Gamma_\theta' \Gamma_\theta = I_{p-1}$ et le vecteur $S_t(\theta) = \Gamma_\theta' X_t / \|\Gamma_\theta' X_t\|$ représente la composante tangentielle de $X_t$. Dans ce cas là, $S_t(\theta)$ est le **signe multivarié**. On peut donc adapter la définition des runs au cadre directionnel avec :

$$R_{1,d}^{(n)} := \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} S_t(\theta)' S_{t-1}(\theta)$$

Dans cet exposé, on s'intéresse au cas où les données sont à symétrie rotationnelle autour de la direction $\theta$. Cela implique que la densité de l'échantillon est de la forme :

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} c_{p,g} g(x'\theta)$$

où $g$ est appelée la fonction angulaire, $c_{p,g}$ est la constante de normalisation. Dans ce contexte, on connait la densité des composantes projetées $v_t(\theta)$, on sait que les composantes tangentielles sont uniformes sur la sphère $\mathcal{S}^{p-2}$ et enfin que les deux composantes sont indépendantes.

## 1.2 Test d'hypothèse : randomness contre dépendance sérielle tangentielle

Dans notre travail, on définit la distribution Markov-tangentielle commme étant, pour un paramètre $\lambda > 0$ :

$$S_1(\theta), \dots, S_n(\theta) \text{ ont pour densité } (s_1, \dots, s_n) \mapsto c_\lambda^n \exp\left(\lambda \sum_{t=2}^{n} s_t' s_{t-1}\right) \text{ sur } (\mathcal{S}^{p-2})^n.$$

Dans ce cas :

- $S_t(\theta) \sim \mathcal{U}_{\mathcal{S}^{p-2}}$,
- $S_t(\theta)|S_{t-1}(\theta) = s_{t-1} \sim VMF(s_{t-1}, \lambda)$ où $\lambda > 0$ est le paramètre de concentration, $s_{t-1}$ est le paramètre de localisation de la loi de Von-Mises Fisher.

**Théorème 1** *Supposons que les composantes $(v_t(\theta))_{t=1,\dots,n}$ sont i.i.d avec pour densité $\tilde{g}$ sur $[-1, 1]$ et sont indépendantes des signes $(S_t(\theta))_{t=1,\dots,n}$ distribués suivant la distribution de Markov-Tangentielle. Alors, vec$(X_1, \dots, X_n)$ a pour densité :*

$$vec(x_1, ..., x_n) \mapsto c_{p,g}^n c_\lambda^n \exp\left(\lambda \sum_{t=2}^{n} s_t' s_{t-1}\right) \prod_{t=1}^{n} g(v_t(\theta))$$

*selon la mesure de surface sur $\mathcal{S}^{p-1}$.On note dans ce cas : vec$(X_1, \dots, X_n) \sim P_{\theta,g,\lambda}$*

Notre travail consiste alors à étudier le test :

"iidness" contre "dépendance sérielle"

Ce qui revient à tester, grâce au Théorème 1 :

$$H_0 : "\lambda = 0" \text{ contre } H_1 : "\lambda > 0"$$

## 1.3 Résultats Principaux

On a montré les résultats suivant :

**Théorème 2** *Sous l'hypothèse nulle, dans le cas où les $S_1(\theta), ..., S_n(\theta)$ sont bien iidness, et considérant la quantité*

$$s_n(\theta) = tr\left[\left(\frac{1}{n}\sum_{t=1}^{n} S_t(\theta)S_t(\theta)'\right)^2\right],$$

*Nous avons :*

$$s_n(\theta)^{-\frac{1}{2}} R_{1,n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Nous étudions également ce qu'il se passe sous l'alternative : on considère une perturbation de l'hypothèse nulle $\left(0+\frac{1}{\sqrt{n}}\ell_n, \theta+\frac{1}{\sqrt{n}}\tau_n\right)$ avec deux suites bornées $(\ell_n)$ et $(\tau_n)$ où $(\tau_n)$ converge vers $0 \neq \tau \in \mathbb{R}^p$ et $(l_n)$ converge vers $0 \neq l \in \mathbb{R}$. On considère le ration log-vraisemblance :

$$\Lambda_n := \log \frac{dP^{(n)}_{\theta+\frac{1}{\sqrt{n}}\tau_n, g, 0+\frac{1}{\sqrt{n}}\ell_n}}{dP^{(n)}_{\theta,g,0}}$$

**Théorème 3 (Résultat LAN)** *Soit $u_n := (\ell_n, \tau_n')'$, on a :*

$$\Lambda_n = u_n'\Delta_n - \frac{1}{2}u_n'\Gamma u_n + o_{\mathrm{P}}(1)$$

*pour $n \to +\infty$ sous $P^{(n)}_{\theta,0,g}$, où $\Delta_{\theta,n} := n^{-1/2}\sum_{t=1}^{n}\varphi_g(v_t(\theta)(1-v_t(\theta)^2)^{1/2}S_t(\theta)$ et $\Delta_{\lambda,n} := n^{-1/2}\sum_{t=2}^{n} S_t(\theta)'S_{t-1}(\theta)$, la suite centrale $\Delta_n := (\Delta_{\lambda,n}, (\Delta_{\theta,n})')'$ est asymptotiquement gausienne (toujours sous $\mathrm{P}^{(n)}_{\theta,0,g}$) de moyenne nulle et pour matrice de covariance*

$$\Gamma := \mathrm{diag}((p-1)^{-1}, \tilde{\Gamma}).$$

De ce résultat LAN, on en déduit le théorème suivant :

**Théorème 4** *D'après le résultat LAN précédent, pour le test :*

$$H_0 : "\lambda = 0" \text{ contre } H_1 : "\lambda > 0"$$

*Le test asymptotiquement et localement plus puissant est le test $\phi^{(n)}_{opt}$ qui rejette l'hypothèse nulle au niveau $\alpha \in ]0,1[$ quand :*

$$\Delta_{\lambda,n}\sqrt{p-1} > z_{1-\alpha}$$

*où $z_{1-\alpha}$ est le quantile de la loi normale.*

Ainsi, nous pouvons construire le test le plus puissant pour l'étude de ce problème avec comme alternative une distribution de Markov-Tangentielle. On en déduit aussi de ce résultat qu'il n'y a pas de coût asymptotique en remplaçant $\theta$ par une estimation tant qu'on utilise un estimateur $\hat{\theta}$ qui est $\sqrt{n}$-consistant. On définit également une statistique de test afin de pouvoir détecter des dépendances au-delà d'un lag de 1.

**Théorème 5** *On définit le run de lag $h \in \mathbb{N}^*$ par :*

$$R_{h,n}(\theta) := \frac{1}{\sqrt{n-h}} \sum_{t=h+1}^{n} S_t(\theta)' S_{t-h}(\theta).$$

*Un test pouvant détecter une dépendance à l'ordre $H \in \mathbb{N}^*$ peut être construit à l'aidre de la statistique*

$$s_n^{-1}(\theta) \sum_{h=1}^{H} (R_{h,n}(\theta))^2.$$

*Nous avons que $s_n^{-1}(\theta) \sum_{h=1}^{H} (R_{h,n}(\theta))^2$ converge en loi vers*

*(i) un chi-deux avec $H$ degrés de liberté sous $\mathrm{P}_{\theta,g}^{(n)}$ et*

*(ii) un chi-deux avec $H$ degrés de liberté et un paramètre de décentralité $(p-1)^{-1}\ell^2$ sous $\mathrm{P}_{\theta,n^{-1/2}\ell_n,g}$, où $\ell := \lim_{n\to+\infty} \ell_n$.*

### 1.4 Simulations Monte-Carlo

Nous appuyons notre travail avec des simulations Monte-Carlo où nous calculons les courbes estimées des puissances de notre test en comparaison avec les tests traditionnels.

Ces courbes (Figure 1) illustrent très bien la propriété de non-coût asymptotique dans le remplacement de $\theta$ par une estimation. De plus, on illustre la puissance qui augmente au fur-et-à-mesure que la valeur de $\lambda$ grandit (et donc que l'on s'éloigne de l'hypothèse nulle). On remarque également que l'on obtient bien la valeur nominale $\alpha$ pour $\lambda = 0$ donc sous $H_0$.

## 2 Application aux données solaires

Pour terminer, dans ce travail, nous proposons d'appliquer nos runs à l'étude des positions des tâches solaires pendant les derniers cycles enregistrés (cycle 11 à 24). En Figure 2, on trouve la représentation des positions de ces tâches solaires au cours du temps et sur plusieurs cycles. Le dégradé illustre l'évolution au cours du cycle (rouge en début de cycle et jaune en fin de cycle). Au vu de ces illustrations, on peut supposer que l'hypothèse de symétrie rotationnelle est vérifiée.

Dans la littérature, une dépendance sérielle à déjà était mise en évidence : la loi de Spörer. Elle permet de modéliser la dépendance des observations par leurs latitudes (en

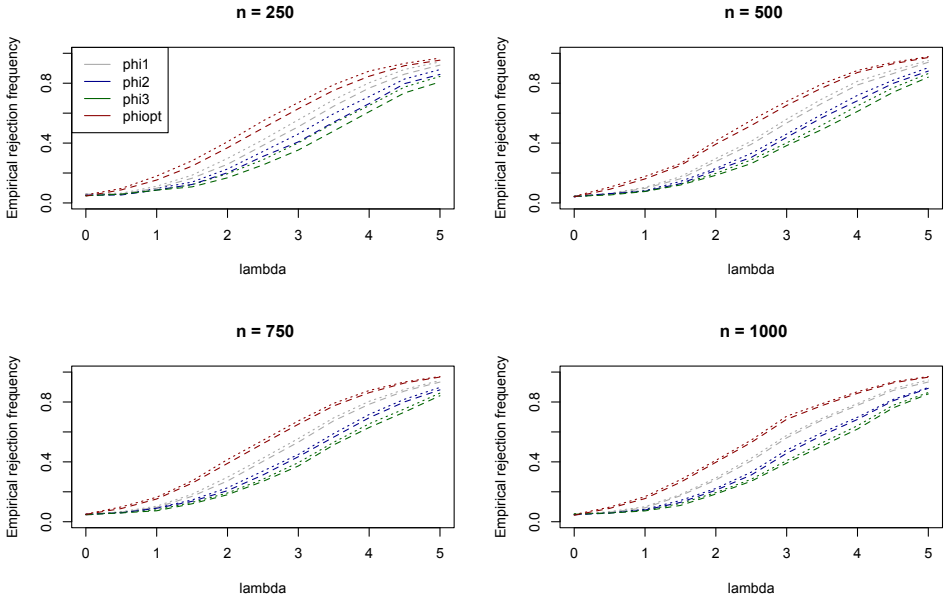



Figure 1: Fréquences de rejets empiriques pour plusieurs tests de runs directionnels : $\phi_1^{(n)}$, $\phi_2^{(n)}$, $\phi_3^{(n)}$ et $\phi_{\text{opt}}^{(n)}$. Avec des pointillés : fréquence de rejet pour un $\theta = (1, 0, 0)'$ alors que la courbe avec des tirets correspond aux calculs de puissances avec un estimateur de $\theta$ et avec $\alpha = 5\%$ pour tous.

valeurs absolues). Avec nos travaux, on souhaite étudier si l'on peut détecter une dépendance également via leurs longitudes. Dans les Figure 3 à 5, on présente les boxplot de p-valeurs pour la partie projetée (latitude) selon l'axe du pôle nord de nos observations (pour chaque cycles), les latitudes en valeurs absolues et les longitudes. Chaque boxplot, pour des lags de 1 à 4, ont été construit sur 200 répétitions de nos tests en conservant à chaque fois un sous échantillon aléatoire représentant 75% de l'échantillon de départ. On en conclut qu'en plus de la dépendance sur les latitudes absolues, une dépendance sérielle le long des longitudes est présente. Le cas du cycle 11 reste encore à expliquer bien que le nombre d'observations est très faible en comparaison des autres cycles.

# Bibliographie

Hentati-Kaffel, R. and De Peretti, P. (2015). Generalized runs tests to detect randomness in hedge funds returns. *Journal of Banking & Finance*, 50:608–615.

Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. *Annals of statistics*, pages 290–298.

Ley, C., Swan, Y., Thiam, B., and Verdebout, T. (2013). Optimal R-estimation of a spherical location. *Statist. Sinica*, 23(1):305–332

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. Wiley Series in Probability and


Figure 2: Positions des tâches solaires pour les cycles 16 à 24.

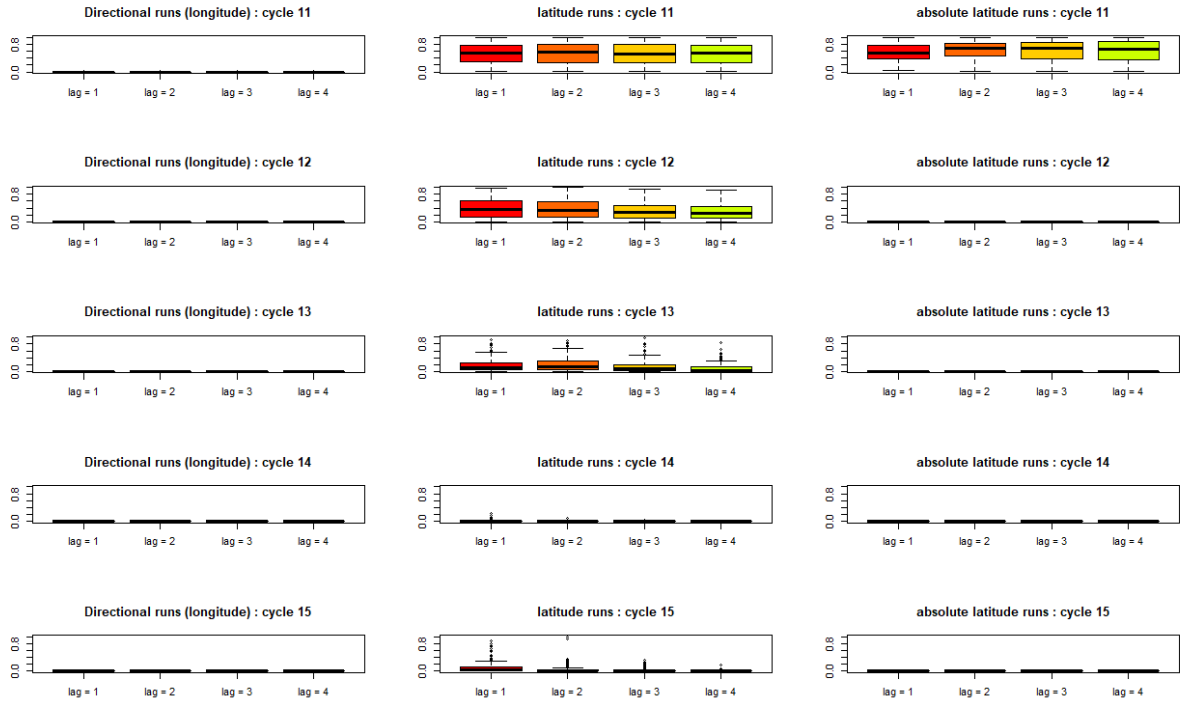



Figure 3: Boxplot des p-valeurs pour les tests de runs pour les latitudes, latitudes absolues et longitudes pour quatre valeurs de lag différentes pour les cycles 11 à 15.

Statistics. Wiley, Chichester.

Paindaveine, D. (2009). On multivariate runs tests for randomness. *Journal of the American Statistical Association*, 104(488):1525–1538.

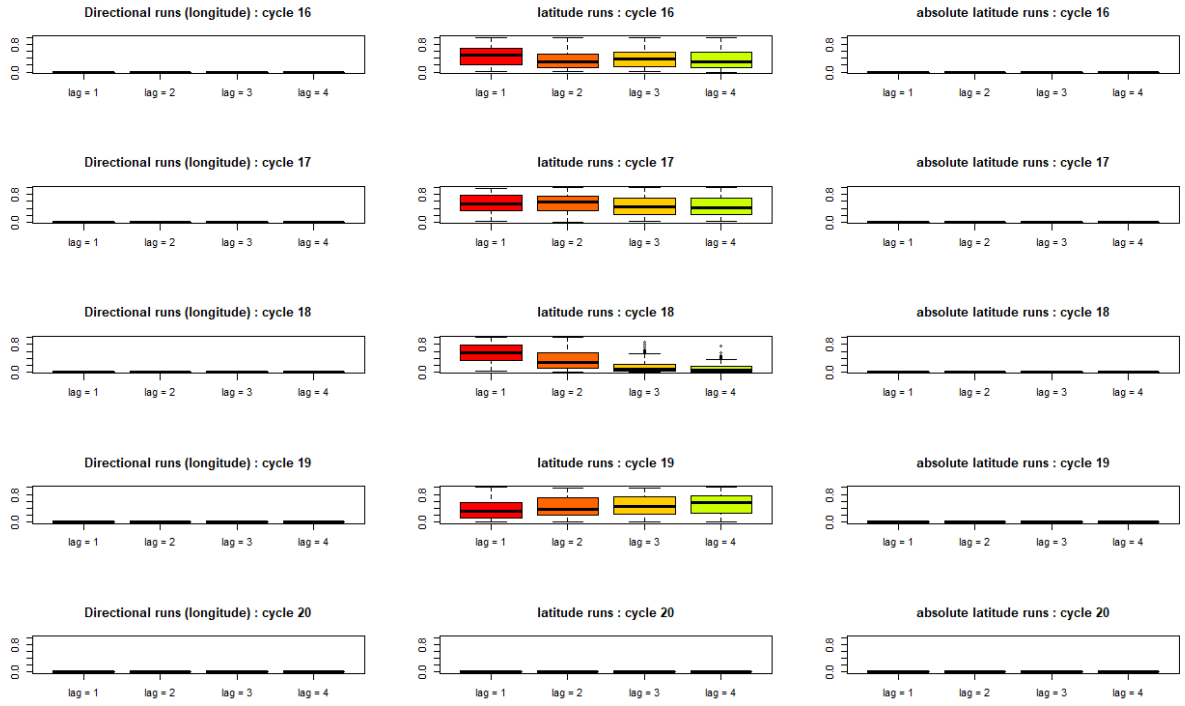



Figure 4: Boxplot des p-valeurs pour les tests de runs pour les latitudes, latitudes absolues et longitudes pour quatre valeurs de lag différentes pour les cycles 16 à 20.

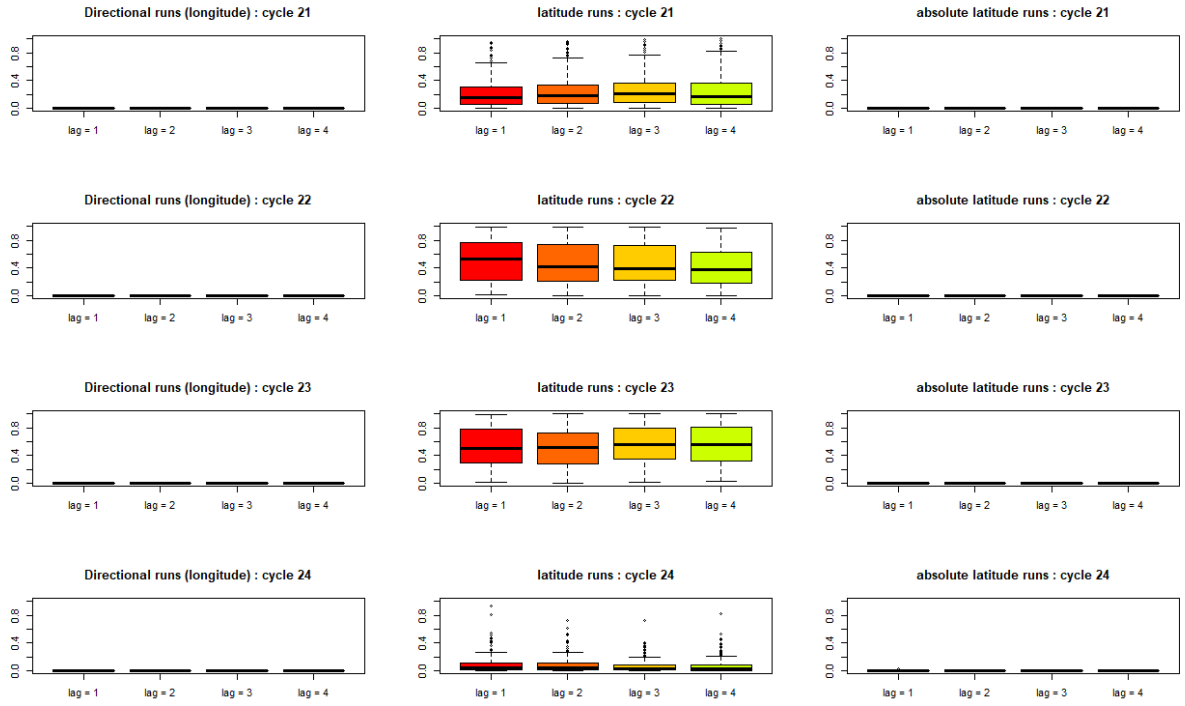



Figure 5: Boxplot des p-valeurs pour les tests de runs pour les latitudes, latitudes absolues et longitudes pour quatre valeurs de lag différentes pour les cycles 21 à 24.

# Processus de Markov déterministes par morceaux

# Deep reinforcement learning for controlled piecewise deterministic Markov process in cancer treatment follow-up.

Alice Cleynen [1] & Benoîte de Saporta [2] & Orlane Rossini [3] & Régis Sabbadin [4] & Meritxell Vinyals [5]

[1] *John Curtin School of Medical Research, Australian National University, Canberra, ACT, Australia and IMAG, Univ Montpellier, CNRS, Montpellier, France alice.cleynen@umontpellier.fr*
[2] *IMAG, Univ Montpellier, CNRS, Montpellier, France benoite.de-saporta@umontpellier.fr*
[3] *IMAG, Univ Montpellier, CNRS, Montpellier, France orlane.rossini@umontpellier.fr*
[4] *Univ Toulouse, INRAE-MIAT, Toulouse, France regis.sabbadin@inrae.fr*
[5] *Univ Toulouse, INRAE-MIAT, Toulouse, France meritxell.vinyals@inrae.fr*

**Résumé.** Les maladies humaines telles que le cancer impliquent un suivi à long terme. Un·e patient·e alterne des phases de rémission et de rechutes. Un biomarqueur est monitoré tout au long du suivi. Sa dynamique est modélisée par un processus de Markov déterministe par morceaux (PDMP) caché et contrôlé. Le PDMP évolue en temps et en espace continus, le processus est observé à travers un bruit et le modèle est partiellement connu,ce qui rend le problème du contrôle particulièrement difficile. À notre connaissance, il n'existe pas de méthode pour contrôler un tel PDMP, c'est-à-dire pour maximiser la vie du·de la patient·e tout en minimisant le coût du traitement et les effets secondaires. Nous considérons des dates discrètes uniquement pour les décisions, transformant ainsi le PDMP contrôlé en un processus de décision markovien partiellement observé (POMDP). L'algorithme deep Q-network (DQN) permet de résoudre le problème de contrôle. Une des limitation de DQN est de ne pas prendre en compte l'historique complet des observations, ce qui est pourtant une caractéristique clé des POMDP. Ce constat nous conduit à traduire le POMDP en un MDP défini sur l'espace des historiques et à appliquer l'algorithme DQN à ce nouveau modèle. Par le biais de simulations, nous comparons les deux méthodes de résolution. Ces analyses visent à éclairer les avantages et les limites de chaque approche dans le contexte du contrôle de PDMP pour une gestion optimale des maladies chroniques.

**Mots-clés.** Processus markovien déterministe par morceaux, états cachés, processus de décision markovien, contrôle stochastique, apprentissage par renforcement profond, optimisation de traitement

**Abstract.** Human diseases such as cancer involve long-term follow-up. A patient alternates between phases of remission with relapses. A biomarker is monitored throughout the follow-up. Its dynamic is modelled by a controlled piecewise deterministic Markov process (PDMP). The PDMP evolves in continuous time and space, the process is observed through noise and some of its parameters are unknown, making the control problem especially difficult. To our knowledge, there is no method to control such a PDMP, i.e. to maximize the life of the patient while minimizing the treatment cost and side effects. We consider discrete

dates only for the decisions, thus turning the controlled PDMP into a partially observable Markov decision process (POMDP). The deep Q-network (DQN) algorithm solves the control problem. A constraint associated with DQN is its inability to consider the entire historical sequence of observations, a crucial aspect in the context of POMDPs. This drawback led us to translate the POMDP into an MDP defined on the space of histories and to apply the DQN algorithm to this new model. Through simulation, we compare the two resolution methods. These analyses aim to shed light on the advantages and limitations of each approach in the context of POMDP control for optimal chronic disease management.

**Keywords.** Piecewise deterministic Markov process, hidden state, Markov decision process, stochastic control, deep reinforcement learning, treatment optimisation

# 1 Introduction

Numerous challenges can be characterized as problems of sequential decision-making under uncertainty, including medical treatment design [Wu+23]. In the field of medical decision-making, the treatment of cancer patients emerges as an intricate challenge. Physicians aim to adapt treatments to uphold the patient's quality of life and life expectancy over time. The primary objective is to formulate optimal strategies for cancer treatment follow-up, acknowledging the continuous nature of the patient's state and its partial observability.

Our focus centres on the computational resolution of a specific category of impulse control problems for piecewise deterministic Markov processes (PDMPs). Impulse control for PDMPs involves selecting actions and intervention dates, as initially explored in [CD89]. Approximating solutions to continuous-time and continuous-state impulse control problems, when the process is only partially observed, jump times remain hidden and the underlying model is partially unknown, presents a challenge. Previous approaches [CS18; CS23], propose to express the controlled PDMP into a partially observable Markov decision process (POMDP). Then they resort to discretizing the state space and employing dynamic programming to approximate the value function effectively addressing problems of continuous state space and partial observability. While effective, these methods are constrained by their reliance on explicit model knowledge and the discretization process. An alternative strategy [Cle+24], adopts a simulation-based approach similar to the partially observable Monte-Carlo planning (POMCP) algorithm [SV10]. This method was essentially developed to deal with the continuous state problem. However, it does not require explicit model information.

In this paper, we propose a resolution method leveraging neural networks. While offering generalization capabilities, our approach aims to approximate the value function directly from a simulator of data. As in previous work, we transform the controlled continuous-time PDMP problem into a discrete-time POMDP. While conventional POMDP solutions often operate over history, deep learning methods frequently focus only on current observations. Hence, a compelling direction emerges in adapting POMDPs to Markov decision processes (MDPs) over history. Our primary conjecture is that this paradigm shift will yield enhanced decision-making policies, optimizing cancer treatment strategies.

The paper is organized as follows. In section 2 we state our optimization problem and turn it into a POMDP. In section 3 we give our resolution strategy and our main assumption. Numerical experiments are also described whereas numerical results are postponed to the upcoming conference.

# 2 Problem statement

In our illustrative medical scenario, a patient enrols in a clinical trial at the onset of a remission phase. Throughout remission, the biomarker hovers at the nominal threshold $\zeta_0$. In the absence of treatment, a relapse triggers an exponential surge in the biomarker level, culminating in the patient's death upon reaching the critical value of $D$. Treatment interventions succeed in lowering the biomarker level, yet with each relapse, the probability of treatment resistance escalates. This intricate interplay involving phases of remission, relapse, and treatment response constitutes the fundamental essence of our impulse control problem. Our investigation starts with delineating a specialized class of impulse control problems designed for piecewise deterministic Markov processes (PDMPs). We describe the translation of our control problem into a partially observable Markov decision process (POMDP) framework.

## 2.1 PDMP

We consider an impulse control problem for hidden piecewise deterministic Markov processes (PDMPs.) We introduce four variables $m, k, \zeta, u$ where the mode $(m, k)$ corresponds to the patient's overall state of health ($m = 0$: remission, $m = 1$: relapse, $m = 2$: untreatable relapse, $m = 3$: death) and $k \in \mathbb{N}$ (the number of curable relapses). The biological marker level is denoted by $\zeta \in [\zeta_0, D]$ with $\zeta_0$ the nominal value and $D$ the death level and $u \in [0, H]$ is the sojourn time in a health' state (added for technical reasons to deal with semi-Markov condition), where $H$ corresponds to the end of the patient's follow-up. The complete state of the patient is denoted by $x = (m, k, \zeta, u)$ in $E$ the state space. Let the state space $E$ be an open subset of $\mathbb{R}^4$ such that : $E \subset \{0, 1, 2\} \times \mathbb{N} \times [\zeta_0, D] \times [0, H] \cup \{3\}$.

Decisions are made throughout a patient's trajectory. Let $\mathbb{D}$ be the space of decisions such that $\mathbb{D} = \mathcal{L} \times \mathcal{R} \cup \{\Delta\}$. Control is expressed as a decision pair: $d = (\ell, r)$, where $r \in \mathcal{R} = \{15, 30, 60\}$ is the delay before the next visit. Visits correspond to the measurement of the biomarker level and the adjustment of the treatment according to results. The therapeutic choice is $\ell \in \mathcal{L} = \{\emptyset, a, b\}$ ($\ell = \emptyset$: *no treatment*, $\ell = a$: *chemotherapy* and $\ell = b$: *palliative care*). The decision $d = \Delta$ corresponds to the action *do nothing* and applies when the patient is dead.

A PDMP on the state space $E$ is defined by three local characteristics $(\mathbf{\Phi}, \lambda, \mathcal{Q})$. The flow $\mathbf{\Phi}$ describes the deterministic trajectory of the process between jumps. The jump intensity $\lambda$ characterizes the frequency of jumps. The Markov kernel $Q$ provides a probabilistic mapping from the pre-jump state to the post-jump state.

The flow depends on the control applied and in particular on the treatment: $\mathbf{\Phi}^{\ell}(x,t) = (m, k, \Phi^{\ell}_{m,k}(\zeta,t), u+t)$, where $\Phi^{\ell}_{m,k}(\zeta,t)$ describes only the trajectory of the biological marker between jumps. When the patient is dead, no treatment is applied and the flow is $\mathbf{\Phi}^{\Delta}(x,t) = (m)$. The biomarker evolution (summarized in Table 1) depends on the therapy choice, the disease regimen and the number of relapses.

Let $t^{\ell\star}(x)$ be the deterministic time the flow takes to reach the boundary of the state space $E$. Let $\partial E = \{1,2\} \times \mathbb{N} \times \{\zeta_0, D\} \times (0, H]$ be the boundary on $E$. The time $t^{\ell\star}(x)$ also depends on the treatment and the disease regimen: $t^{\ell\star}_{m,k}(\zeta) = \inf\{t > 0 : \Phi^{\ell}_{m,k}(\zeta,t) \in \partial E\}$. This function is detailed in table 2.

| $m/\ell$ | $\emptyset$ | $a$ | $b$ |
|---|---|---|---|
| 0 | $\zeta_0$ | | |
| 1 | $\zeta e^{v_1 t}$ | $\zeta e^{-\frac{v_1}{k}t}$ | $\zeta e^{v_1 t}$ |
| 2 | $\zeta e^{v_2 t}$ | | |

Table 1: **Flow.** $\Phi^{\ell}_{m,k}(\zeta,t)$, where $v_1$ and $v_2$ are constants.

| m/$\ell$ | $\emptyset$ | $a$ | $b$ |
|---|---|---|---|
| 0 | $+\infty$ | | |
| 1 | $\frac{1}{v_1}\log(\frac{D}{\zeta})$ | $\frac{k}{v_1}\log(\frac{\zeta}{\zeta_0})$ | $\frac{1}{v_1}\log(\frac{D}{\zeta})$ |
| 2 | $\frac{1}{v_2}\log(\frac{D}{\zeta})$ | | |

Table 2: **Boundary jump.** $t^{\ell\star}_{m,k}(\zeta)$,, where $v_1$ and $v_2$ are constants.

Treatment also influences the risk function $\lambda^{\ell}(x) = \lambda^{\ell}_{m,k}(\zeta,u)$. Notably, there are two distinctive types of relapse scenarios considered: standard relapses occurring during remission phases and relapses indicative of therapeutic escape. For standard relapses, the probability of occurrence increases with the duration of time spent in remission. On the other hand, the risk of relapses associated with therapeutic escape is influenced by the biomarker level. In light of these considerations, we choose Weibull distributions of the form: $\mu_i(u) = (\alpha_i u)^{\beta_i}$ and $\mu'_2(\zeta) = (\alpha'\zeta)^{\beta'}$. Details of jump intensity are available in table 3.

| m / $\ell$ | $\emptyset$ | $a$ | $b$ |
|---|---|---|---|
| 0 | $(\mu_1+\mu_2)(u)$ | $\mu_2(u)$ | $(\mu_1+\mu_2)(u)$ |
| 1 | $\mu'_2(\zeta)$ | | |
| 2 | 0 | | |
| 3 | 0 | | |

Table 3: **Jump intensity.** $\lambda^{\ell}_{m,k}(x)$

In remission, the patient may transition to either a curable relapse in the absence of chemotherapy or an incurable relapse. In the case of relapse and without treatment, the biomarker increases to a critical value $D$, leading to the patient's death. When chemotherapy is administered, the biomarker decreases to $\zeta_0$ and returns to remission. Regardless of treatment chosen, therapeutic escape may occur at any time. In the case of therapeutic escape, the biomarker increases, regardless of the administered treatment, toward the $D$ threshold, ultimately resulting in the patient's death. We define the Markov kernel $\mathbf{Q}(x,\ell)(x')$ in table 4, for all $h : E \to \mathbb{R}$ a bounded measurable test function. Case $m = 3$ is omitted as no jumps are allowed when patients are dead.

Let $\mathcal{P}(x,d)(x')$ be the transition kernel associated with the continuous time PDMP, for a time period $r$ such that $\mathcal{P}h(x,d) = \mathbb{E}[h(X_r)|X_0 = x, d = (\ell, r)]$. The transition kernel of the PDMP combines the deterministic flow, the jump intensity and the Markov kernel. However,

| | $\ell \in \{\emptyset, b\}$ |
|---|---|
| $m = 0$ | $h(1, k+1, \zeta_0, 0)\frac{\mu_1(u)}{(\mu_1+\mu_2)(u)} + h(2, k, \zeta_0, 0)\frac{\mu_2(u)}{\mu_1+\mu_2)(u)}$ |
| $m = 1$ | $h(2, k, \zeta, 0)\mathbb{1}_{D>\zeta} + h(3)\mathbb{1}_{\zeta=D}$ |
| $m = 2$ | $h(3)\mathbb{1}_{\zeta=D}$ |
| | $\ell = a$ |
| $m = 0$ | $h(2, k, \zeta_0, 0)$ |
| $m = 1$ | $h(2, k, \zeta, 0)\mathbb{1}_{\zeta>\zeta_0} + h(0, k, \zeta_0, 0)\mathbb{1}_{\zeta=\zeta_0}$ |
| $m = 2$ | $h(3)\mathbb{1}_{\zeta=D}$ |

Table 4: **Markov kernel.** $\mathbf{Q}(x, \ell)(x')$

due to its extensive nature, detailed analytic formulas will not be included in this paper, but it is worth noting that they allow the kernel to be simulated easily.

## 2.2 Partially observed Markov decision process

The trajectory of the process defined above depends on the sequence of decisions and the dates on which the decisions are made. The visit dates take place at discrete dates $n_0 = 0, n_1, ..., n_k$, where the time lapse between two visits can be 15, 30 or 60 days. At most, $N = \frac{H}{15}$ visits can occur. The impulse control problem described above can be formalized as a discrete-time partially observed Markov decision process (POMDP).

A POMDP is a tuple $(\mathbb{S}, \Omega, \mathbb{D}, \mathbb{K}, \mathcal{T}, C)$, where $\mathbb{S}$ corresponds to the state space, which corresponds to the PDMP state space $E$, $\Omega$ corresponds to the observation space, $\mathbb{D}$ to the decision space, which remains unchanged, $\mathbb{K}(\omega) \subseteq \Omega \times \mathbb{D}$ is the space of admissible decisions in observation $\omega$, $\mathcal{T}(s, \omega, d)(s', \omega')$ is the transition kernel of a state-observation tuple $(s, \omega) \in \mathbb{S} \times \Omega$ to state-observation tuple $(s', \omega') \in \mathbb{S} \times \Omega$ when action $d \in \mathbb{K}(\omega)$ is taken, $c(s, d)$ is the cost incurred in state $s \in S$ when decision $d \in D$ is made.

Blood measurements are intrinsically subject to variations independent of the medical condition. These fluctuations can be attributed to measurement errors, natural variations, and external influences. The biomarker is thus observed through a multiplicative noise as the biomarker is growing exponentially. Let $y = \zeta e^{\epsilon}$ with $\epsilon \sim \mathcal{N}(0, 1)$ be the noisy biomarker. In addition, the patient's overall health is not observed, except when the patient is deceased. Let $z = \mathbb{1}_{(m=3)}$ be the death indicator. Decision-related constraints then appear. The last visit must take place at the end $H$ of the follow-up. The variable $t \in [0, H]$ indicates the time elapsed since the start of the trajectory. In addition, treatment must be applied for a minimum of 45 days. The variable $\tau \in [0, H]$ corresponds to the time since treatment (chemotherapy or palliative care) was administered. At a given time $t$, the observation of a patient's condition is $\omega = (\tau, t, y, z)$ with $\omega \in \Omega$. The observation space is $\Omega \subset [0, H]^2 \times \mathbb{R}_+ \times \{0\} \cup [0, H] \times \{1\}$.

Let $\mathbb{K} \subseteq \Omega \times \mathbb{D}$, be the constraint space. It is used to specify all allowed actions state by state: $\mathbb{K}(\omega) = \{d \in \mathbb{D}; (\omega, d) \in \mathbb{K}\} \neq \emptyset$. Constraints are only defined by observations.

$$\mathbb{K}(\omega) = \begin{cases} \{\Delta\} & \text{if } z = 1 \text{ or } t = H \\ (l, r) \in \{a, b\} \times \mathcal{R} & \text{if } 0 < \tau < 45 \text{ and } t + r \leq H \\ (l, r) \in \mathcal{L} \times \mathcal{R} & \text{such that } t + r \leq H \end{cases}$$

The POMDP joint transition-observation function can be expressed as a function of $\mathcal{P}(x, d)(x')$ the piecewise deterministic Markov process (PDMP) transition kernel. For all $g : \mathbb{S} \times \Omega \to \mathbb{R}$ be a bounded measurable test function, $\mathbb{1}_z(m, k, \zeta, u, \tau, t, y, z) = \mathbb{1}_{z=1}$ and $f_\epsilon$ is the probability density function of $\epsilon$ . Let $\mathcal{T}g(s, \omega, d) = \mathbb{E}[g(S_{t+r}, \omega_{t+r})|S_t = (m, k, \zeta, u), \omega_t = (\tau, t, y, z), d]$.

$$\mathcal{T}g(s, \omega, d) = \begin{cases} g(3, H, 1) & \text{if } d = \Delta \\ \mathcal{P}g(m, k, \zeta, u, 0, t + r, \zeta\epsilon, 0) & \text{if } \ell = \emptyset \\ \mathcal{P}\mathbb{1}_{z=1}g(m, k, \zeta, u, H, 1) + \int \mathcal{P}\mathbb{1}_{z\neq 1}g(m, k, \zeta, u, \tau + r, t + r, \zeta e^{f_\epsilon(\xi)}, 0)\mathrm{d}\xi & \text{else} \end{cases}$$

Let $C$ be the non-negative cost-per-stage function such that $C : \mathbb{D} \times \mathbb{S} \to \mathbb{R}_+$. In POMDPs, the cost function quantifies the cost associated with different decisions per stage.

A history is a sequence of observations and decisions $h_n = \{\omega_0, d_0, \omega_1, \cdots, \omega_n\}$ and $H$ is the set of histories. Along a trajectory, the agent applies decision rules which map a history to an appropriate decision. Let $f_n : H_n \to \mathbb{K}(\omega_n)$ be a decision rule for the kth visit. We define an admissible policy $\pi$ as a sequence of decision rules $\pi = (f_n)_{0:N-1}$ and $\Pi$ the set of all admissible policies. Then, the total cumulated cost from visit $k$ is defined as follows $C_k = \sum_{n=k}^{N-1} C(D_n, S_{n+1})$ for all $h \in H$.

The value function $V^\pi(h_k) = \mathbb{E}_\pi[C_k|h = h_k]$ is the expected return from history $h$ when following policy $\pi$. Our next aim is to obtain an optimal policy $\pi^\star$ such that the value function $V$ is optimal: $V^\star(h) = \min_{\pi\in\Pi} V^\pi(h)$ for all $h \in H$.

# 3 Resolution strategy

In the next section, we proceed to an exploration of the deep Q-network (DQN) algorithm. We then move on to the translation of the partially observable Markov decision process (POMDP) into a Markov decision process (MDP) on the history. This allows us to discuss the two main strategies proposed.

## 3.1 Deep Q-Network algorithm

Reinforcement learning methodologies can be broadly categorized into two principal approaches: value learning and policy learning. These approaches diverge in their strategies for addressing sequential decision problems. Value learning focuses on assessing and enhancing the value function associated with a given policy, aiming to identify the optimal value for each state. On the other hand, policy learning directly updates the policy, determining the optimal sequence of actions for each state. Notably, policy iteration often achieves convergence in

fewer iterations, yet value iteration assures convergence to the optimal policy. Furthermore, the value iteration approach ensures a deterministic policy, a crucial characteristic in cancer monitoring and treatment.

Deep Q-network (DQN) is a value learning algorithm developed in [Mni+13]. DQN uses a deep neural network to approximate the action-state function $Q$ rather than the value function $V$, where the Q-function corresponds to the expected return starting from state $s \in \mathbb{S}$, taking the decision $d \in \mathbb{D}$: $Q^{\pi}(h_k, d_k) = \mathbb{E}_{\pi}[C_k | h = h_k, d = d_k]$. The optimisation problem is then $V^{\star}(h) = min_{d \in \mathbb{D}} Q(h, d)$. This choice aims to mitigate the overestimation of Q-values, thereby contributing to faster and more stable learning during training. DQN facilitates optimal decision-making in intricate and dynamic environments by focusing on the Q function.

The double deep Q-network (DDQN), introduced in [HGS16], represents an enhancement of the DQN. In contrast to DQN, DDQN employs two neural networks: a target network and a primary network, as illustrated in Figure 1. The primary network is responsible for action selection, while the target network is utilized to compute target values for the primary network. The loss function calculation depends on the weights $\theta$ of each network: $L(\theta) = \left[(c + \gamma \max_{d_{t+1}} Q(\omega_{t+1}, d_{t+1}; \theta^{-})) - Q(\omega, d^{\star}; \theta)\right]^2$, where $\gamma \in [0, 1]$ denotes a discount factor. The training objective is to minimize this loss function to improve the predictive capabilities of the neural network. By segregating action selection and Q-value estimation processes, DDQN mitigates Q-value overestimation, resulting in expedited training, improved learning stability, and more effective policies.

DQN has introduced methodologies like experience replay to enhance network updates and model training [Mni+13]. Experience replay involves the utilization of a replay buffer, as depicted in Figure 1. The replay buffer is a repository of past experiences, storing transitions consisting of state-action pairs, the next state, and their corresponding costs. During iterations, a random batch of experiences is sampled from the buffer, diminishing the inherent correlation in sequential data and breaking temporal dependencies between successive observations. This detachment of experiences contributes to a more resilient and stable training procedure, preventing the algorithm from being overly affected by the immediate consequences of recent actions. Consequently, the incorporation of a replay buffer enhances the stability and efficiency of the learning process.

## 3.2 Equivalent MDP on the history

The deep Q-network (DQN) algorithm operates within the framework of a Markov decision process (MDP). Our problem is a partially observable Markov decision process (POMDP). To bridge this gap, a necessary transformation is required to convert our POMDP into an MDP defined in the space of histories. Employing the MDP framework on the histories will empower us to base our decisions on the entire trajectory, as opposed to only relying on the last observation. This modification is anticipated to enhance the performance of the DQN algorithm by providing a more comprehensive context for decision-making within our sequential control problem.

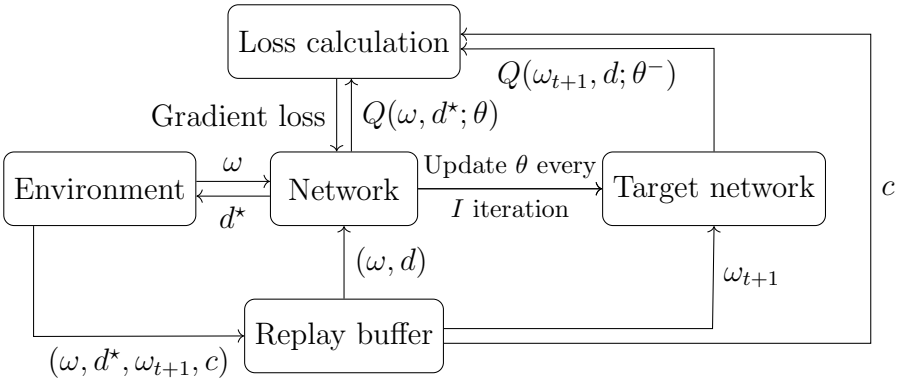



Figure 1: Conceptual diagram of a double deep Q-network. *The primary network estimates the Q-value for a given observation $\omega$. The target network provides an estimate of the Q-value for observation $\omega_{t+1}$, based on the last copy of the weights in the main network. For every $I$ learning iteration, the weights of the main network are updated in the target network.*

Consider the POMDP described in section 2.2 and defined by the tuple $(\mathbb{S}, \mathbb{D}, \mathbb{K}, \mathcal{T}, C)$. Let $\mathcal{B}(s_k, h) = \mathbb{P}(s_k \in S | h_k = h)$ be the belief state, i.e. the probability distribution over states given history $h \in H_k$. It is updated as the agent takes actions and receives observations, allowing it to make decisions based on records of past observations.

Consider the derived MDP with histories as states, defined by the tuple $(H, \mathbb{D}, \mathbb{K}, \tilde{\mathcal{T}}, \tilde{C})$, where $\tilde{\mathcal{T}}(h, d)(h') = \int_{s \in \mathbb{S}} \int_{s' \in \mathbb{S}} g(s') \mathcal{B}(ds, h) \mathcal{T}(s, d)(ds')$ and $\tilde{C}(h, d) = \int_{s \in \mathbb{S}} \mathcal{B}(ds, h) C(s, d)$. The value function $\tilde{V}^\pi(h)$ of the MDP on history is equal to the value function $V^\pi(h)$ of the POMDP, for every $\pi \in \Pi$. The detailed proof is available in [SV10].

## 3.3 Numerical experiments

In our experimental setup, we aim to investigate and compare the performance of two distinct strategies for solving our sequential decision-making problem. The first scenario involves applying the (double) deep Q-network (DQN) algorithm within the partially observable Markov decision process (POMDP) framework, where the algorithm relies only on the current observation for decision-making, as illustrated in Figure 2. In contrast, the second scenario entails leveraging the DQN algorithm within the Markov decision process (MDP) framework. The algorithm takes into account the entire history of observations when making decisions, as depicted in Figure 3. This comparative analysis will provide insights into the impact of historical information on the algorithm's decision-making process and overall performance.

$$\omega_k = (\tau, t, y, z) \longrightarrow \boxed{\text{DQN algorithm}} \longrightarrow \left\{ \begin{array}{l} Q(\omega_k, d_1) \\ \cdots \\ Q(\omega_k, d_n) \end{array} \right\} \longrightarrow d^\star(\omega_k) = argmax_{d_i} Q(\omega_k, d_i)$$

Figure 2: DQN Applied to POMDP

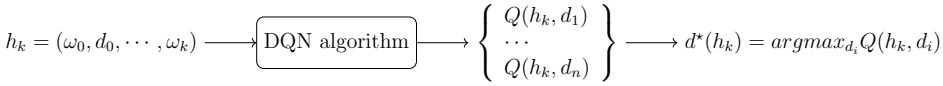



Figure 3: DQN Applied to MDP on history

The efficacy of decision-making policies is evaluated based on their cost implications, with superior policies invariably associated with lower costs. We expect that DQN within the MDP framework will outperform. This hypothesis is grounded in the idea that a richer context, encapsulating the entire history of interactions, can lead to more informed decision-making. The results of our comparative analysis will be presented and discussed at the upcoming conference.

# 4 Conclusion

In conclusion, the monitoring of cancer treatment in patients can be modelled by a hidden controlled piecewise deterministic semi-Markov process (PDsMP). The formalism of this process is complex and does not allow for its direct resolution. For this reason, its transformation into an equivalent partially observable Markov decision process (POMDP) is essential. Typically, POMDPs are solved over the space of histories, yet deep learning methods in RLlib often focus only on observations. By translating the POMDP into a Markov decision process (MDP) over histories we can use deep Q-networks (DQN) to account for the entire historical context. Our underlying hypothesis posits that this approach will yield a more effective policy. The outcomes of this exploration will be presented on the day of the conference.

The exploration of alternative modelling avenues remains a compelling direction for future research. Instead of exclusively adopting the MDP on history, an intriguing avenue could involve transitioning towards an MDP formulated on belief states. This shift could offer a more nuanced representation of uncertainty and enhance decision-making capabilities. Additionally, while our present study focuses on translating a POMDP into an MDP on historical states for integration with DQN, it is crucial to acknowledge existing methods utilizing Recurrent Neural Networks (RNNs) directly on historical sequences [HS15; Kap+18]. Finally, we hypothesize that a more informative framework leads to more efficient decision-making. Consequently, exploring model-based approaches, particularly Bayesian model-based methods for learning the model, presents a promising avenue for future investigations.

# References

[CD89] O. L. V. Costa and M. H. A. Davis. “Impulse control of piecewise-deterministic processes”. en. In: *Mathematics of Control, Signals, and Systems* 2.3 (1989), pp. 187–206. DOI: 10.1007/BF02551384.

[Cle+24] Alice Cleynen et al. “Medical follow-up optimization: A Monte-Carlo planning strategy”. In: *arXiv* (2024). DOI: 10.48550/arXiv.2401.03972.

[CS18] Alice Cleynen and Benoıte de Saporta. “Change-point detection for piecewise deterministic Markov processes”. In: *Automatica* 97 (Nov. 2018), pp. 234–247. DOI: 10.1016/j.automatica.2018.08.011.

[CS23] Alice Cleynen and Benoıte de Saporta. “Numerical method to solve impulse control problems for partially observed piecewise deterministic Markov processes”. In: *arXiv* (2023). DOI: 10.48550/arXiv.2112.09408.

[HGS16] Hado van Hasselt, Arthur Guez, and David Silver. “Deep reinforcement learning with double Q-Learning”. In: *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 2016, pp. 2094–2100. DOI: 10.5555/3016100.3016191.

[HS15] Matthew Hausknecht and Peter Stone. “Deep Recurrent Q-Learning for Partially Observable MDPs”. In: *Papers from the 2015 AAAI Fall Symposium* (July 2015). DOI: 10.48550/arXiv.1507.06527. eprint: 1507.06527.

[Kap+18] Steven Kapturowski et al. “Recurrent Experience Replay in Distributed Reinforcement Learning”. In: *International conference on learning representations* (Sept. 2018).

[Mni+13] Volodymyr Mnih et al. “Playing Atari with Deep Reinforcement Learning”. In: (2013). DOI: 10.48550/arXiv.1312.5602.

[SV10] David Silver and Joel Veness. “Monte-Carlo Planning in Large POMDPs”. In: *Advances in Neural Information Processing Systems*. Vol. 23. Curran Associates, Inc., 2010.

[Wu+23] XiaoDan Wu et al. “A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis”. en. In: *npj Digital Medicine* 6.1 (2023), pp. 1–12. DOI: 10.1038/s41746-023-00755-5.

# Sojourn time estimation in partially observed piecewise deterministic Markov processes — application to myeloma modeling

Alice Cleynen[1,2] & Benoîte de Saporta[1] & Amélie Vernay[1]

[1] *Univ Montpellier, CNRS, Montpellier, France, alice.cleynen@umontpellier.fr, benoite.de-saporta@umontpellier.fr, amelie.vernay@umontpellier.fr*
[2] *John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia*

**Résumé.** Nous considérons un problème d'estimation du temps de rechute dans des Processus de Markov Déterministes par Morceaux (PDMP) dont la composante euclidienne est un biomarqueur de substitution pour l'état de patients atteints de myélome. L'une des principales difficultés du problème réside dans le fait que notre processus n'est que partiellement observé, ce que peu de travaux ont pris en compte jusqu'à présent. Nous proposons une méthode basée sur de la régression itérative pour estimer les paramètres d'un PDMP observé en temps discret et avec du bruit. Nous évaluons les performances de notre procédure à travers une étude de simulation et discutons des limites de notre approche.

**Mots-clés.** Processus de Markov Déterministes par Morceaux, Temps de rechute, Observations partielles, Survie

**Abstract.** We consider a problem of relapse time estimation in Piecewise Deterministic Markov Processes (PDMPs) whose Euclidean component is a proxy biomarker for the status of myeloma patients. One of the main difficulties of the problem lies in the fact that our process is only partially observed, which few works have considered until now. We provide a method based on iterative regression for estimating the parameters of a PDMP observed in discrete time and through noise. We assess the performances of our procedure through a simulation study and discuss the limitations of our approach.

**Keywords.** Piecewise Deterministic Markov Processes, Relapse time, Partial observations, Survival

## 1 Introduction

When monitoring patients suffering from a disease, it is common to measure the level of a specific biomarker as an indicator of the pathological process or the action of a treatment. The biomarker may evolve in a deterministic manner during different phases of the disease and be disrupted by random jumps that indicate a change in the patient's condition. In this work, we propose to model this behaviour through Piecewise Deterministic Markov Processes (PDMPs) (Davis, 1984) in the case of patients followed after developing myeloma. Our aim is to estimate patients' relapse time, or survival time, based on their biomarker levels measured during follow-up visits. One of the difficulties of the problem lies in the fact that our observations are noisy and only partially observed: we have a measurement of a biomarker at discrete visit dates, but its true value at and between dates is unknown. Such frameworks are considered e.g. in (Cleynen and de Saporta, 2018) where the authors propose a numerical scheme to approximate the value function of a

change-point detection problem for a partially observed PDMP in discrete time and through noise. In this paper, we focus on relapse time estimation. We develop an estimation procedure to recover the parameters of our model and evaluate its performance through a simulation study. We discuss the impact of parameter estimation errors on the overall relapse time estimation.

# 2 The model

## 2.1 PDMPs

Piecewise Deterministic Markov Processes (PDMPs) introduced by Davis in the 80's are a general class of stochastic processes, including almost all non-diffusion models found in applied probability (Davis, 1984). These continuous-time processes are used to describe deterministic motions punctuated by random jumps.

Let $(X_t)_{t\geq 0}$ be a PDMP defined on a state space $E \subset \mathbb{R}^d$. The trajectories of $(X_t)_{t\geq 0}$ are determined by the behavior of the process between jumps, as well as when and where the jumps occur. These aspects are described by a flow $\Phi$, a jump intensity $\lambda$ and a Markov kernel $Q$, respectively. The flow $\Phi\colon E \times \mathbb{R}_+ \to E$ is a continuous function satisfying the semi-group property: $\forall x \in E,\ \forall t, s \in \mathbb{R}_+,\ \Phi(x, t+s) = \Phi(\Phi(x,t), s)$. Starting from $x \in E$, $\Phi(x,t)$ gives the position of the process after some time $t$ if no jump has occurred (see Figure 1).

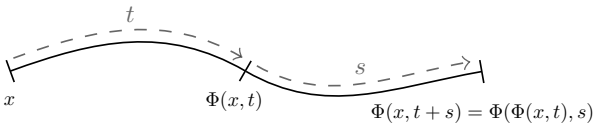



Figure 1 – Starting from $x$ at time 0, the process follows its flow up to time $t$ and ends up at $\Phi(x,t)$, assuming no jump has occurred. Going on up to time $t+s$ is the same as starting from $\Phi(x,t)$ and following the flow for a time $s$.

The process can jump deterministically or randomly. Deterministic jumps occur when the flow reaches the boundary $\partial E$ of $E$. Given a starting point $x \in E$, this happens after a time $t^*(x) = \inf\{t > 0\colon \Phi(x,t) \in \partial E\}$. Random jumps are governed by the jump intensity $\lambda\colon E \to \mathbb{R}_+$ — also known as the hazard rate — which is a measurable function such that $\forall x \in E,\ \exists \varepsilon > 0\colon \int_0^\varepsilon \lambda(\Phi(x,s))ds < \infty$. That is, jumps cannot occur instantaneously (and therefore there cannot be several jumps at the same time). The jump times of a PDMP are obtained by taking the minimum between deterministic jumps and stochastic ones. Given a starting point $x_0 \in E$, for all $t \in \mathbb{R}_+$, the first jump time $T_1$ satisfies

$$\mathbb{P}_{X_0=x_0}(T_1 > t) = \begin{cases} \mathbb{P}_{X_0=x_0}(T_1 > t) = \mathrm{e}^{-\int_0^t \lambda(\Phi(x_0,s))ds} & \text{if} \quad t < t^*(x_0) \\ 0 & \text{if} \quad t \geq t^*(x_0) \end{cases}. \tag{1}$$

For both deterministic and random jumps, the new location of the PDMP is drawn from the Markov kernel $Q\colon \bar{E} \times \mathcal{B}(\bar{E}) \to E$, where $\mathcal{B}(\bar{E})$ is the set of Borels of $\bar{E}$. When the process starts from $x \in \bar{E}$, we have that $\forall A \in \mathcal{B}(\bar{E}),\ Q(x,A) = \mathbb{P}(X_{T_1} \in A \mid X_{T_1^-} = x)$, where $T_1^-$ denotes the time just before the first jump. For $Q$ to be a Markov kernel, we need $x \mapsto Q(x,A)$ to be measurable $\forall A \in \mathcal{B}(\bar{E})$ and $A \mapsto Q(x,A)$ to be a probability density function $\forall x \in \bar{E}$. The Markov kernel $Q$ satisfies $\mathbb{P}(X_t = x \mid X_{t^-} = x) = 0, \forall t \in \mathbb{R}_+$. In other words, each jump must involve a real change of location.

It is common practice to separate the state space $E$ into a hybrid one made up of a discrete component and a continuous one, such that $X_t = (m_t, \zeta_t) \in E \subset M \times \mathbb{R}^d$, where $m_t$ corresponds to a discrete mode and $\zeta_t$ to a continuous variable. Furthermore, the state space can be specific to each mode: $\forall m \in M,\ E_m \subset \mathbb{R}^{d_m}$. The mode-specific flow $\Phi_m$ is such that $\forall m \in M,\ \Phi_m\colon E_m \times \mathbb{R}_+ \to E_m$ and $\Phi((m,\zeta),t) = (m, \Phi_m(\zeta,t))$.

### 2.2 Our model

We consider subjects undergoing medical follow-up after developing myeloma. During medical visits, their serum-M protein levels, a proxy for the progression of their disease, are measured. At the start of monitoring they are administered a treatment, the effect of which is to reduce the serum-M protein level exponentially. If the level falls below a certain fixed threshold $\zeta_0$, the patient is considered to be in remission. In the event of a relapse, the serum level rises again exponentially. The horizon $H$ of follow-up is different for each patient. We now link the notations introduced in 2.1 with our specific model. There are three possible modes for the subjects in the study. They can either be sick under treatment ($m = -1$), sick without treatment ($m = 1$) or in remission ($m = 0$) — for simplicity, we assume that the process characteristics in remission mode are the same with or without treatment and do not differentiate the two cases. We thus have $M = \{-1, 0, 1\}$ and $E_{-1} = ]\zeta_0, +\infty[$, $E_0 = \{\zeta_0\} \times \mathbb{R}_+$ and $E_1 = [\zeta_0, +\infty[$. In mode $m = 0$, a time variable $u \in \mathbb{R}_+$ is added to the state space to allow more flexibility in the jump intensity while ensuring the Markov property holds. Under treatment, the serum spike decreases exponentially with a slope $v_{-1} < 0$. During relapse, it increases exponentially with a slope $v_1 > 0$. For all $\zeta \in \mathbb{R}$ and for all $u, t \in \mathbb{R}_+$ we have

$$\begin{cases} \Phi_{-1}(\zeta, t) = \zeta \mathrm{e}^{v_{-1}t}, \\ \Phi_0(\zeta, t, u) = \zeta = \zeta_0, \\ \Phi_1(\zeta, t) = \zeta \mathrm{e}^{v_1 t}. \end{cases} \tag{2}$$

In mode $m = -1$, the jump occurs when the subject reaches the fixed remission threshold $\zeta_0$. This is therefore a deterministic jump at the boundary and $\lambda_{-1}(\zeta) = 0$. The jump time $t^*_{-1}(\zeta)$ is the solution of $\Phi_{-1}(\zeta, t) = \zeta_0$. That is, $t^*_{-1}(\zeta) = \frac{1}{v_{-1}} \log(\frac{\zeta_0}{\zeta})$. With the additional time variable in mode $m = 0$, we have $\Phi_0((\zeta_0, u, t), t) = (\zeta_0, u + t)$ and $t^*_0(\zeta, u) = +\infty$. The jump intensity $\lambda_0 > 0$ is unknown. For practical reasons explained in Section 4, we decide to approximate it with a Weibull distribution. In our model, we consider that once the process reached mode $m = 1$, no more jump can occur.

For the mode-specific Markov kernels, we have

$$\begin{cases} Q_{-1}(m', \zeta', u' \mid \zeta) = \mathbb{1}_{\zeta=\zeta_0} \times \mathbb{1}_{\zeta'=\zeta_0} \times \mathbb{1}_{m'=0} \times \mathbb{1}_{u'=0} \\ Q_0(m', \zeta' \mid \zeta, u) = \mathbb{1}_{\zeta=\zeta_0=\zeta'} \times \mathbb{1}_{m'=1} \end{cases} \tag{3}$$

## 3 Real data

Our data comes from a study carried out by the Inter-Groupe Francophone du Myélome (IFM) in 2009. The author consider the effect of lenalidomide, bortezomib and dexamethasone (RVD) therapy alone versus RVD therapy plus autologous stem cell transplantation on disease progression (Attal et al., 2017). About 700 patients with newly diagnosed myeloma were randomly divided into two groups and followed up after receiving their respective therapy. Their serum M-protein levels were measured at different frequencies depending on the phase of the trial. Patients may remain in remission, suffer a relapse or leave the study for various reasons. The length of follow-up therefore varies from one individual to another. In this work, we are interested in estimating the relapse time of patients. For the time being, for the sake of simplicity we do not take into account the difference between patient groups, nor any other covariate, and we consider relapse times in a general way. The raw data had been preprocessed to remove observations unsuitable for model fitting. The detailed preprocessing procedure is given in Appendix A. The post-processed dataset consists of 479 serum M protein peak trajectories over time.

# 4 Estimating model parameters

In this section, we explain the estimation procedure for the parameters of our model based on the data. This involves first estimating the process jump times, and then finding the parameters of the relapse time distribution. Note that our observations here are not in continuous time, so the jump times are hidden (see Figure 2). In what follows, we let $t_0 := T_1$ the time of the first jump, from mode $m = -1$ to mode $m = 0$ and $T_0 := T_2$ the second jump, if any, from mode $m = 0$ to mode $m = 1$.

## 4.1 Jump time estimation

Our estimation method for $t_0$ and $T_0$ is an iterative optimisation process based on regression. It is illustrated in Figure 2. Let $(X_t)_{t\geq 0}$ be a PDMP as defined in 2.2 and let $(X_k)_{k\in\mathbb{N}} = (X_{d_k})_{k\in\mathbb{N}}$ be the process at the observation dates $(d_k)_{k\in\mathbb{N}}$ that generates our trajectories. The observations are defined as $Y_k = F(X_k)\mathrm{e}^{\varepsilon_k}$, where $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise and where $F\colon E \to \mathbb{R}_+$ is a function that returns the second component of the PDMP. Hence, $Y_k = \zeta_k \mathrm{e}^{\varepsilon_k}$. We use a multiplicative noise both to match the exponential growth and decay of biomarker level and to simplify the estimation procedure described hereafter. We start with the estimation of $t_0$. Let $y_0, y_1, \ldots, y_j$ be the first $j$ values of an $N$-length trajectory of M-protein levels recorded at dates $d_0, d_1, \ldots d_j$, respectively. The biomarker level has an exponential form, so we use least squares to fit a linear function $f\colon x \mapsto ax + b$ to the logarithm of our data, where $a^* = \widehat{v_{-1}}$ and $\mathrm{e}^{b^*} = \widehat{x_0}$ are the optimal solutions of the problem. We use a logarithmic transformation to prevent errors at the beginning of the trajectory from having too much weight on the overall error.

We then estimate $t_0$ as the solution for $t$ of $\widehat{x_0}\mathrm{e}^{\widehat{v_{-1}}t} = \zeta_0$ (see Figure 2). This gives us an approximation of the jump time from $m = -1$ to $m = 0$ and we compute a general regression error as the sum of two errors: one between the points falling before $\widehat{t_0}$ and the fitted curve, and the other on the remaining part of the trajectory. Note that all the above estimates depend on $j$, which is omitted for clarity. This process is repeated for $j \in \{3, \ldots, N\}$ until a stopping criterion is met. This results in optimal estimates for the entry time into remission $t_0$ and the slope $v_{-1}$ in mode $m = -1$. Note that $\widehat{v_{-1}}$ is only used to estimate $t_0$ and will not be used to estimate the relapse time afterwards. Details of the estimation procedure can be found in Algorithm 1.

We can then use the same process again on the remaining part of the trajectory[1] — that is, on $(y_k)_{k=\widehat{t_0},\ldots,N}$ — to obtain $\widehat{T_0}$ and $\widehat{v_1}$. It is important to note that if the trajectory is very flat, it is not relevant to look for a jump time, plus the algorithmic minimization could fail due to numerical instability. In such cases, we assume that no change in mode occurred. This happens mainly when subjects do not relapse within their follow up time. They are considered "censored subjects" and are discussed hereafter.

[1] From a computational point of view, if $k_{\widehat{t_0}}$ is the index of time $\widehat{t_0}$ in the vector $d \in \mathbb{R}^N$ of visit dates associated with the trajectory, we apply Algorithm 1 with $d_{k_{\widehat{t_0}}:N}$ in reverse order of its elements as vector of visit dates. Same for the corresponding vector of spikes.

**Algorithm 1:** Jump time estimation

**input:** $d \in \mathbb{R}^N$ vector of visit dates, $y \in \mathbb{R}^N$ vector of spikes at visit dates, $\zeta_0 \in \mathbb{R}$ theoretical threshold for remission mode

**init:** $\Delta_{\text{tmp}} = \Delta_{\min} = \infty$

**for** $j = 3, \dots, N$ **do**

- $d_{\text{tmp}} = d_{0:j}$ `//slicing of the first` $j$ `coordinates of` $d$
- $y_{\text{tmp}} = y_{0:j}$
- find $\widehat{a}$ and $\widehat{b}$ optimal solutions when fitting $f : x \mapsto ax + b$ to $\log(y_{\text{tmp}})$ using least squares
- $t_{0\text{tmp}} = (\log(\widehat{\zeta_0})) - \log(\widehat{b})/\widehat{a}$ `//solve` $\widehat{b}e^{t\widehat{a}} = \zeta_0$ `for` $t$
- $k = |i : d_i \leq t_{0\text{tmp}}|$
- $n_1 = \|y_{0:\lfloor k \rfloor} - \widehat{b}e^{-d_{0:\lfloor k \rfloor} \times \widehat{a}}\|_2^2$ `//error between the first` $\lfloor k \rfloor$ `points and the fitted curve`
- $n_2 = \|y_{\lfloor j \rfloor + 1:N} - \zeta_0\|_2^2$ `//error on the remaining part of the trajectory`
- $\Delta = \Delta_{\text{tmp}}$
- $\Delta_{\text{tmp}} = n_1 + n_2$
- **if** $\Delta_{\text{tmp}} \leq \Delta_{\min}$ AND $t_{0\text{tmp}} > 0$ **then**
  - $t_0 = t_{0\text{tmp}}$
  - $\Delta_{\min} = \Delta_{\text{tmp}}$
  - $a^* = \widehat{a}$
  - $b^* = \widehat{b}$
- **if** $\Delta_{\text{tmp}} > \Delta$ AND $t_0 > 0$ AND $j > 15$ **then** break

**return** $t_0, a^*, b^*$

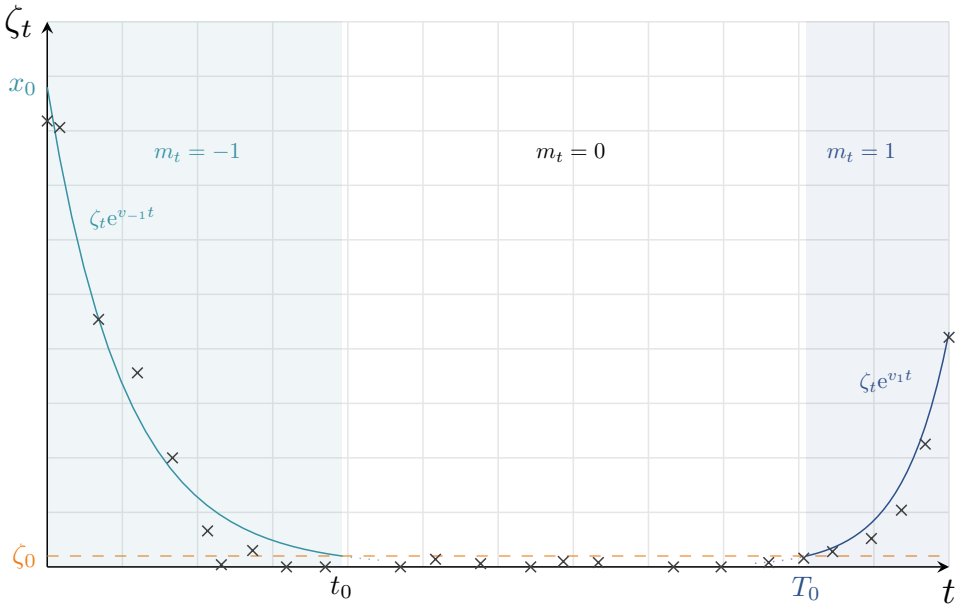



Figure 2 – Illustration of the estimation process. The true PDMP starts in mode $m = -1$ from an initial point $x_0$ and follows a deterministic trajectory along its flow unit the first jump occurs at time $t_0$. The mode changes to become $m = 0$ and the flow equals $\zeta_0$ until a new jump occurs at time $T_0$. The mode switches to $m = 1$ and the trajectory rises exponentially along $\Phi_1(\zeta, t)$. The black crosses represents the observations and we seek to recover the model parameters.

### 4.2 Survival time before relapse

Having calculated $\widehat{t_0}$ and $\widehat{T_0}$ for every trajectory, we now have access to the survival times of subjects. That is, the number of days between the remission start and the beginning of the relapse, if any, or the follow up time otherwise. Following the terms of survival analysis, an *event* is defined as the occurrence of a relapse. A patient is considered *censored* if no event has occurred until the end of its follow up. The survival function $S(t) = \mathbb{P}(T_0 - t_0 > t)$ gives us the probability that a patient remain in remission beyond a time $t$ after remission entry, and we seek to recover its parameters. We assume that the hazard rate — the event rate at time $t$ conditional on survival up to time $t$ or later — increases with time. This leads us to choose a Weibull distribution to model our relapse time, as is conventionally done in such cases in survival analysis. We will see in Section 5 that this assumption is quite reasonable. The probability density function of the Weibull distribution is given by

$$f(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} \mathrm{e}^{-\left(\frac{x}{\beta}\right)^{\alpha}}, \tag{4}$$

for $x > 0$ and where $\alpha$ is a shape parameter and $\beta$ is a scale parameter. Its hazard function is

$$h(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1}. \tag{5}$$

Note that a shape parameter $\alpha < 1$ (resp. $\alpha > 1$) means that the failure rate decreases (resp. increases) over time. If $\alpha = 1$, this rate is constant. To estimate the parameters of the Weibull distribution, we fit a parametric survival regression model with the survival time as a response variable together with an censoring indicator. This gives us the estimations $\widehat{\alpha}$ and $\widehat{\beta}$ of the shape and scale parameters, respectively.

## 5 Simulation study

In this section, we assess the performances of the estimation procedure described in Section 4 on simulated data. The trajectories are generated from the model presented in 2.2. We compare the estimates with the ground-truth and discuss the impact of jump times errors on relapse time estimation. The parametric survival regressions are performed with the `R survival` package (Therneau, 2023).

### 5.1 Simulation process

We use Algorithm 2 to simulate patient trajectories. The input parameters are chosen as follows. When patients enter the clinical trial, they can have a serum M-protein level anywhere between a remission threshold and some upper threshold with equiprobability. This leads us to opt for a uniform distribution to generate the first point $x_0 = \zeta_{t=0}$ of a trajectory. Similarly, patients follow-up may end at any time within a certain date range since they did not enter the study at the same time. We thus choose $H$ uniformly distributed as well. Based on real data analysis, we pick $l_0 = 15$, $u_0 = 55$ and $l_H = 900$, $u_H = 1900$ as lower and upper bounds for $\zeta_{t=0}$ and $H$, respectively. To select the model parameters $\alpha$, $\beta$, $v_{-1}$ and $v_1$ for the simulation, we apply the estimation procedure on our real data. We take $v_{-1}$ to be the average of all the estimated slopes in mode $m = -1$. The same is done for $v_1$ but only considering trajectories for which a relapse is predicted. This gives us $v_{-1} = -0.046$ and $v_1 = 0.012$. The parametric survival regression gives $\widehat{\alpha} = 4.69$ and $\widehat{\beta} = 1650$ which we choose as our $\alpha$ and $\beta$ inputs. Note that $\widehat{\alpha} > 1$, which is consistent with the fact that the risk of a relapse increases over time. Figure 3 shows the fitted survival curve and the overlapping Weibull survival function plotted with $\widehat{\alpha}$ and $\widehat{\beta}$. The Weibull distribution seems to fit the data fairly well, although its survival curve is slightly shifted compared to the survival curve adjusted on the data. Both the shape and the scale estimates are obtained with significant p-values ($2 \times 10^{-16}$). On clinical trials, visit frequency often varies during follow-up depending on the stage of the study and the patient status. For the sake of

simplicity, we only consider fixed time intervals $\delta \in \mathbb{N}$ between visits. However, we study several different values for $\delta$ to assess the impact of visit frequency on estimation errors. Finding an optimal frequency is an important point: visits must be frequent enough to detect a relapse as early as possible, while avoiding a burdensome and restrictive monitoring for patients. We consider values of $\delta$ in $\{10, 20, 30, 40, 50, 60\}$. The remission threshold $\zeta_0$ is set to 1 according to medical criteria. For practical reasons, we use additive noise to simulate our data, although we have assumed in 4.1 that the noise is multiplicative. This provides trajectories that more closely resemble those of the real data. We set the standard deviation $\sigma$ to 1.

**Algorithm 2:** Simulation of one trajectory from a PDMP

**input:** $l_0, u_0 \in \mathbb{R}$, $l_H, u_H \in \mathbb{R}$, lower and upper bounds for starting point and follow up time distribution, $\alpha, \beta \in \mathbb{R}$ shape and scale parameters for the Weibull distribution, $v_{-1}, v_1 \in \mathbb{R}$ slopes for mode $m = -1$ and $m = 1$, $\delta \in \mathbb{N}$ number of days between two visit dates, $\zeta_0 \in \mathbb{R}$ theoretical threshold for remission mode

**init:** $\zeta_{t=0} \sim \mathcal{U}_{[l_0,u_0]}$, $H \sim \mathcal{U}_{[l_H,u_H]}$

$t_0 \leftarrow (\log(\zeta_0) - \log(\zeta_{t=0}))/v_{-1}$ ; $w \sim \mathcal{W}(\alpha, \beta)$

$T_0 \leftarrow w + t_0$ ; $c \leftarrow \mathbb{1}_{\{H \leq T_0\}}$ //c censoring indicator

$\delta_{\text{end}} \leftarrow \lfloor H/\delta \rfloor$ //last visit date

**for** $k = 0, \ldots, \delta_{\text{end}}$ **do**

  $d_k \leftarrow k\delta$ //visit dates at regular time intervals until $H$

  $m_k \leftarrow -\mathbb{1}_{\{d_k < t_0\}} + \mathbb{1}_{\{d_k > T_0\}}\mathbb{1}_{\{c\}}$

  **if** $m_k = -1$ **then**

    $\zeta_k \leftarrow \Phi_{-1}(\zeta_0, d_k)$

  **else if** $m_k = 0$ **then**

    $\zeta_k \leftarrow \zeta_0$

  **else**

    $\zeta_k \leftarrow \Phi_1(\zeta_0, d_k - T_0)$

  $y_k = \zeta_k + \mathcal{N}(0, \sigma^2)$

**return** $y = (y_k)_k$

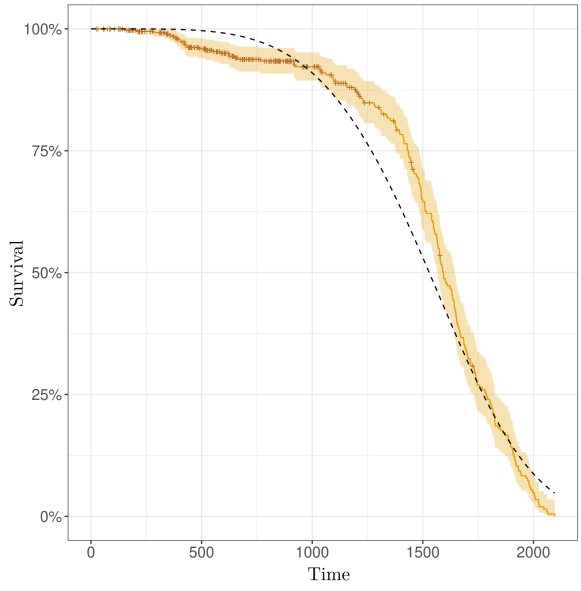



Figure 3 – Survival curve fitted on real data and 95% confidence interval. Crosses signify censored events. The black dashed line shows the Weibull survival function with parameters estimated from survival regression.

## 5.2 Results

We evaluate our method through 100 repetitions of 500 PDMP trajectories.

Figure 4 presents the distributions of absolute errors on $\widehat{t_0}$ and $\widehat{T_0}$ depending on visit frequency. Unsurprisingly, with longer time intervals between visits, the mean error and the variability increase for both jump times. For $\delta \leq 30$, the mean error on $\widehat{t_0}$ is less that the one on $\widehat{T_0}$, whereas the opposite occurs for larger values of $\delta$. Note that for $\delta = 60$, the average number of days of error is approximately equal to the time interval itself: the estimate $\widehat{T_0}$ is one visit apart from the actual jump time $T_0$. As a complement to Figure 4, Table 1 presents the mean number days of error on relapse time estimation $|(T_0 - t_0) - (\widehat{T_0} - \widehat{t_0})|$ depending on visit frequency. One can see that the errors on $\widehat{t_0}$ and $\widehat{T_0}$ do not compensate: roughly speaking, the mean error on relapse time estimation is the sum of the two, especially for larger values of $\delta$. Figure 5 shows the distributions of relative errors on Weibull shape and scale parameter estimates. The average error on the scale parameter $\beta$ increases with the time between visits. This is fairly consistent with the results in Table 1, since the increase in $\delta$ degrades the estimate of relapse time and spreads out its distribution, leading to a less accurate estimation. Conversely, the average relative error on $\widehat{\alpha}$ decreases with $\delta$, but this behaviour is less clear to us to interpret. However, it can be said that a compromise needs to be found on the frequency of visits for the estimation of the distribution parameters. The curves of the probability density functions of the Weibull distributions obtained with the mean estimates are shown in Figure 6, together with the ground-truth density used to simulate the trajectories. The modes of the distributions are closer to the true mode for smaller values of $\delta$. However, the estimated distributions are less spread out than the true one, whatever the visit frequency. On the whole, we tend to underestimate the relapse time. Note however that of all the trajectories simulated, about 66% were censored, which adds a significant difficulty to the problem. Table 2 provides the average proportions of false censoring and false relapse predictions. A false censoring occurs when the $T_0$ estimates falls after the time horizon $H$. Such errors tend to appear more often as the visit frequency decreases: the longer we wait before checking a patient again, the more likely we are to miss a relapse. Our estimation procedure never predicts a relapse when there is none. This is probably due to the low noise level we have chosen for the simulations, and it would be interesting to consider noisier trajectories to see if this behaviour is maintained.

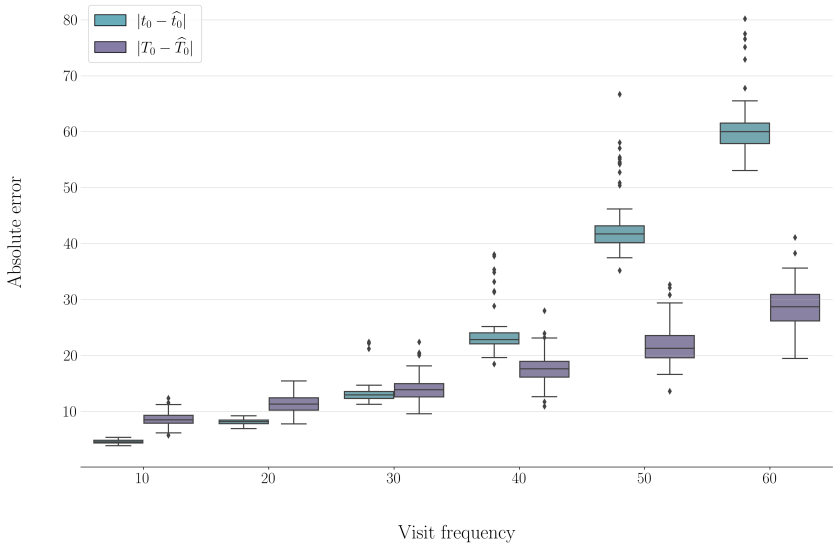


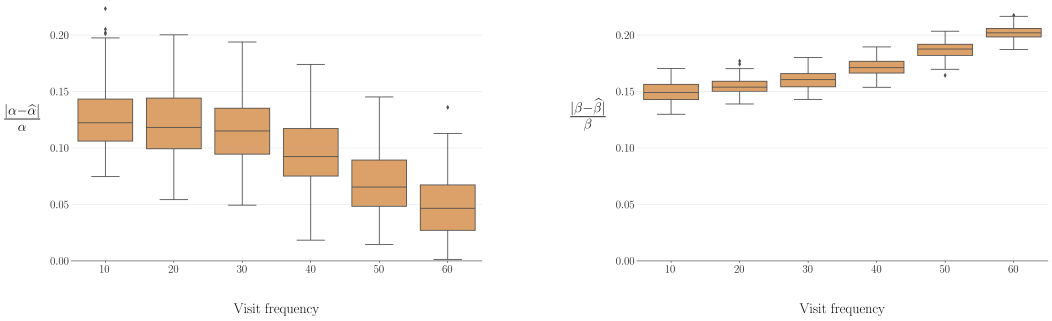

Figure 4 – Boxplots of absolute number of days of error on estimated jump times $\widehat{t_0}$ and $\widehat{T_0}$ over 100 repetitions with 500 trajectory samples, depending on visit frequency (in days). The distributions are only calculated on trajectories for which relapse occurred and is correctly predicted.

| Visit frequency | $\lvert(T_0 - t_0) - (\widehat{T_0} - \widehat{t_0})\rvert$ |
|---|---|
| 10 | 10.745± 1.156 |
| 20 | 16.494± 1.738 |
| 30 | 24.429± 2.931 |
| 40 | 38.650± 4.552 |
| 50 | 62.341± 6.471 |
| 60 | 86.816± 5.793 |

Table 1 – Average absolute number of days error and standard deviation on estimated survival times over 100 repetitions with 500 trajectory samples, depending on visit frequency. The mean is only calculated on trajectories for which relapse occurred and is correctly predicted. Values are rounded to the nearest $10^{-3}$.

| Visit frequency | False censoring | False relapse |
|---|---|---|
| 10 | 0.092 | 0.0 |
| 20 | 0.096 | 0.0 |
| 30 | 0.101 | 0.0 |
| 40 | 0.109 | 0.0 |
| 50 | 0.124 | 0.0 |
| 60 | 0.144 | 0.0 |

Table 2 – Average proportion of false censoring and false relapse prediction over 100 repetitions with 500 trajectory samples, depending on visit frequency. Values are rounded to the nearest $10^{-3}$.




(a) (b)

Figure 5 – Boxplots of relative error on Weibull (a) shape and (b) scale parameter estimates over 100 repetitions with 500 trajectory samples, depending on visit frequency (in days).

## 6 Conclusion and further work

We have presented a proof of concept for the estimation of relapse time in PDMPs with noisy and partially observed trajectories. The results are encouraging, but there is still considerable room for improvement. For the moment, we have only considered a simplistic simulation framework and it would seem appropriate to choose the initial parameters differently and to control the censoring rate. We have chosen to fully exploit the assumptions of our model. It would be interesting to see what happens if one or more of them are no longer verified. What would happen if our noise were additive rather than multiplicative? Or if the growth and decay of the process were not exponential? We could also consider the case where the slopes are no longer fixed but chosen at random. Finally, it would be interesting to compare our estimation procedure with methods for which no or almost no assumptions about the model are required, such as moving average methods or hidden Markov models. This comparison is planned as future work.

## 7 Acknowledgement

We acknowledge the support of European Union's Horizon 2020 research and innovation program (https://marie-sklodowska-curie-actions.ec.europa.eu, Marie Sklodowska-Curie grant agreement No 890462, to Alice Cleynen), and of the French National Research Agency (ANR), under grant ANR-20-CHIA-0001 (https://anr.fr/, project CAMELOT, to Amélie Vernay).

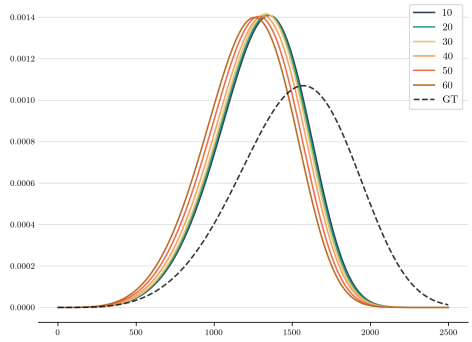



Figure 6 – Curves of the Weibull probability density functions with average parameters estimates over 100 repetitions with 500 trajectory samples, depending on visit frequency (in days). The dashed curve shows the ground-truth density.

# Bibliography

Michel Attal, Valerie Lauwers-Cances, Cyrille Hulin, Xavier Leleu, Denis Caillot, Martine Escoffre, Bertrand Arnulf, Margaret Macro, Karim Belhadj, Laurent Garderet, et al. Lenalidomide, bortezomib, and dexamethasone with transplantation for myeloma. *New England Journal of Medicine*, 376(14):1311–1320, 2017.

Alice Cleynen and Benoîte de Saporta. Change-point detection for piecewise deterministic markov processes. *Automatica*, 97:234–247, 2018. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2018.08.011. URL https://www.sciencedirect.com/science/article/pii/S0005109818304011.

MHA. Davis. Piecewise-deterministic Markov processes: a general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B*, 46(3):353–388, 1984. With discussion.

Terry M Therneau. *A Package for Survival Analysis in R*, 2023. URL https://CRAN.R-project.org/package=survival. R package version 3.5-7.

# A Preprocessing real data

The raw data had been preprocessed to remove observations unsuitable for model fitting. Apart from observations for which the quantity of interest is missing or incomplete, the following data have been deleted, in this order:

1. first observations whose spike is lower than that of the second; the process is repeated iteratively until the first spike is higher than the next one
2. trajectories whose first spike is lower than a threshold of 5µg/L (value at which the level is considered negligible)
3. trajectories with less than two spikes below 5µg/L: we assume that the associated subjects never reached remission
4. observations with spike equalling 0µg/L surrounded by spikes above 5µg/L: we assume these correspond to insignificant isolated zeros
5. trajectories with less than 10 visits.

# Échantillonnage préférentiel dynamique informé par des graphes

Guillaume Chennetier [1,2] & Hassane Chraibi [1] & Anne Dutfoy [1] & Josselin Garnier [2]

[1] *EDF Lab Paris-Saclay, France.*
{guillaume.chennetier,hassane.chraibi,anne.dutfoy}@edf.fr
[2] *CMAP, CNRS, Ecole polytechnique, Institut Polytechnique de Paris, France.*
{guillaume.chennetier,josselin.garnier}@polytechnique.edu

**Résumé.** L'échantillonnage préférentiel est l'une des méthodes de réduction de variance les plus populaires pour l'estimation de quantités de la forme $\mathbb{E}_{X\sim\mathbf{p}}[\varphi(X)]$. Déterminer une distribution d'importance efficace est cependant notoirement délicat en grande dimension. Un cas typique de grande dimension, pourtant relativement peu traité, est celui où $X$ ne représente pas un vecteur mais la trajectoire d'un processus stochastique. Nous avons en tête l'évaluation de la fiabilité de systèmes industriels complexes dont le fonctionnement est modélisé par des processus stochastiques. Nous proposons une nouvelle famille de distributions d'importance adaptées à la simulation d'événements rares pour des processus de Markov non-diffusifs, c'est-à-dire des processus de Markov déterministes par morceaux (abrégés PDMPs). Ces processus évoluent selon des équations différentielles déterministes dont les paramètres sont soumis à des sauts aléatoires. La distribution d'importance optimale pour ces PDMPs est caractérisée par la fonction dite "committor" du processus. Celle-ci associe à toute trajectoire partielle du PDMP, la probabilité que la trajectoire complète réalise l'événement d'intérêt sachant son passé. Notre méthodologie s'articule en trois phases. On approxime d'abord notre PDMP par une marche aléatoire homogène sur un graphe pour laquelle on peut calculer explicitement des temps moyens d'atteinte. On construit ensuite une famille d'approximations de la fonction committor à partir de ces temps d'atteinte. Enfin, on détermine séquentiellement un bon candidat au sein de cette famille (et par conséquent une densité d'importance efficace) en minimisant un critère d'entropie croisée.

**Mots-clés.** Monte-Carlo, Échantillonnage préférentiel adaptatif, Simulation d'événements rares, Processus de Markov déterministes par morceaux, Marche aléatoire sur graphes.

**Abstract.** Importance sampling is one of the most popular variance reduction methods for estimating quantities of the form $\mathbb{E}_{X\sim\mathbf{p}}[\varphi(X)]$. However, determining an effective importance distribution is notoriously challenging in high dimensions. A typical case of high dimensionality, yet relatively underexplored, is when $X$ does not represent a vector but the trajectory of a stochastic process. We have in mind the reliability assessment of complex industrial systems whose operation is modeled by stochastic processes. We propose a new family of importance distributions tailored for rare event simulation with any non-diffusive Markov processes, i.e., any piecewise deterministic Markov processes (abbreviated PDMPs). These processes evolve according to deterministic differential equations whose parameters are subject to random jumps. The optimal importance distribution for these PDMPs is characterized by the so-called "committor" function of the process. This function associates with

any partial trajectory of the PDMP the probability that the complete trajectory realizes the event of interest, given its past. Our methodology consists of three phases: first, we approximate our PDMP by a homogeneous random walk on a graph for which mean hitting times can be explicitly computed; then, we construct a family of committor function approximations based on these hitting times; finally, we sequentially determine a good candidate within this family (and consequently an effective importance density) by minimizing a cross-entropy criterion.

**Keywords.** Monte Carlo, Adaptive Importance Sampling, Rare event simulation, Piecewise deterministic Markov processes, Random walk on graphs.

# 1 Cadre et limites des méthodes de Monte-Carlo

On se place dans le cadre de l'inférence par simulation, où l'on tente d'estimer des quantités de la forme $\overline{\varphi} := \mathbb{E}_{X\sim\mathbf{p}}[\varphi(X)]$ avec $\mathbf{p}$ pour distribution de référence et $\varphi$ une fonction d'intérêt. Une méthode de Monte-Carlo est alors l'approche la plus naturelle pour estimer une telle quantité : elle consiste à approcher $\overline{\varphi}$ par une moyenne empirique $\frac{1}{n}\sum_{k=1}^{n}\varphi(X_k)$ à partir d'un échantillon $X_1,\dots,X_n$ i.i.d. que l'on a généré numériquement sous la distribution de référence $\mathbf{p}$. Cette méthode est simple à mettre en oeuvre à condition de pouvoir échantillonner $\mathbf{p}$ à coût raisonnable. Malheureusement, elle se révèle inefficace dès que la distribution $\mathbf{p}$ ne met pas assez de poids là où la fonction d'intérêt $\varphi$ prend de grandes valeurs (en valeur absolue). La variance de l'estimateur Monte-Carlo devient alors trop grande pour estimer $\varphi$ avec un budget restreint. Un cas typique qui nous intéresse est celui de la simulation d'événement rare où $\varphi(X) = \mathbb{1}_{X\in\mathbf{F}}$ avec $\mathbf{F}$ un ensemble rarement visité sous la distribution $\mathbf{p}$.

# 2 Échantillonnage préférentiel

L'échantillonnage préférentiel est l'une des méthodes de réduction de variance les plus connues et peut-être la plus directe. Elle consiste à générer un échantillon $X_1,\dots,X_n$ sous une distribution alternative $\mathbf{g}$ (dite "distribution d'importance") et à retourner la même moyenne empirique mais cette fois-ci pondérée par le rapport de vraisemblance approprié. On suppose d'une part que $\mathbf{p}$ et $\mathbf{g}$ sont deux densités de probabilités sur un espace $\mathcal{X}$ dominées par une mesure $\mu$, et d'autre part que le support de $\mathbf{g}$ est inclus dans le support de $\varphi\times\mathbf{p}$.

$$\begin{aligned}\overline{\varphi} := \mathbb{E}_{X\sim\mathbf{p}}[\varphi(X)] &= \int_{\mathcal{X}}\varphi(x)\mathbf{p}(x)\mu(\mathrm{d}x) = \int_{\mathcal{X}}\varphi(x)\frac{\mathbf{p}(x)}{\mathbf{g}(x)}\mathbf{g}(x)\mu(\mathrm{d}x)\\ &= \mathbb{E}_{X\sim\mathbf{g}}\left[\varphi(X)\frac{\mathbf{p}(X)}{\mathbf{g}(X)}\right] \approx \frac{1}{n}\sum_{k=1}^{n}\varphi(X_k)\frac{\mathbf{p}(X_k)}{\mathbf{g}(X_k)} \quad \text{avec } X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \mathbf{g}. \qquad (1)\end{aligned}$$

La distribution d'importance optimale est caractérisée par la densité $\mathbf{g}_{\text{opt}} \propto |\varphi| \times \mathbf{p}$. Elle produit un estimateur de variance nulle lorsque la fonction d'intérêt $\varphi$ est de signe constant. Cette distribution est bien entendu inaccessible mais elle guide l'objectif en pratique : générer un échantillon $X_1, \ldots, X_n$ sous une distribution $\mathbf{g}$ la plus proche possible de $\mathbf{g}_{\text{opt}}$ pour obtenir un estimateur de $\overline{\varphi}$ de faible variance.

En dehors de cas jouets simples, cet objectif est délicat. On a le plus souvent recours à des méthodes dites adaptatives, où la distribution d'importance est raffinée au cours des simulations. L'une des plus connues, appelée méthode d'entropie croisée [1], consiste à déterminer la distribution d'importance $\mathbf{g} \in \mathcal{G}$ la plus proche de la distribution optimale $\mathbf{g}_{\text{opt}}$ au sens de la divergence de Kullback-Leibler.

$$\begin{aligned} \arg\min_{\mathbf{g}\in\mathcal{G}} \mathcal{D}_{\text{KL}}\left(\mathbf{g}_{\text{opt}} \| \mathbf{g}\right) &= \arg\min_{\mathbf{g}\in\mathcal{G}} \mathbb{E}_{\mathbf{g}_{\text{opt}}}\left[\log\left(\frac{\mathbf{g}_{\text{opt}}(X)}{\mathbf{g}(X)}\right)\right] \\ &= \arg\max_{\mathbf{g}\in\mathcal{G}} \mathbb{E}_{\mathbf{p}}\left[|\varphi(X)| \log\left(\mathbf{g}(X)\right)\right]. \end{aligned} \tag{2}$$

Le problème reste notoirement difficile en grande dimension. D'une part car les rapports de vraisemblance "dégénèrent" : ils tendent à prendre des valeurs de plus en plus proches de zéro lorsque la dimension augmente mais restent d'espérance $\mathbb{E}_{\mathbf{g}}\left[\frac{\mathbf{p}(X)}{\mathbf{g}(X)}\right] = 1$. Et d'autre part, la méthode adaptative est confronté au problème d'estimation de densité, très sensible à la dimension de l'espace. Différentes approches ont été proposées pour contourner le problème telles que la réduction de dimension par projection dans des sous-espaces bien choisis [2, 3], ou l'usage de méthodes d'apprentissage modernes que l'on retrouve derrière les modèles génératifs [4].

# 3 Processus de Markov déterministes par morceaux

La dimension de l'espace ne pose pas uniquement problème lorsque $X$ est un vecteur de grande taille mais aussi lorsqu'il s'agit de la trajectoire d'un processus stochastique. Nous avons typiquement en tête l'évaluation de la fiabilité de systèmes industriels dynamiques dont le fonctionnement est modélisé par des processus stochastiques (voir notre article précédent [5]). On se place dans le cadre général des processus de Markov non-diffusifs qui forment la classe des processus de Markov déterministes par morceaux (abrégés PDMPs).

Un PDMP est un processus stochastique dit hybride car sa variable d'état $X_t = (Z_t, V_t) \in \mathcal{X} = (\mathcal{Z} \times \mathcal{V})$ est constituée d'une partie continue $Z_t$ appelée position et d'une partie discrète $V_t$ appelée régime. La position du PDMP évolue selon un système d'équations différentielles déterministes paramétré par le régime. Lorsque la position atteint la frontière de l'espace d'état, ainsi qu'à des instants aléatoires, le processus saute vers un nouvel état lui aussi choisi aléatoirement. La distribution des instants de sauts et de la destination des sauts dépend du régime et continûment de la position. Introduits par Mark HA Davis dans les années 1980 [6], le lecteur intéressé trouvera une introduction moderne aux PDMPs orientée fiabilité industrielle, dans [7].

On s'intéresse à la probabilité qu'une trajectoire de PDMP $(X_t)_{t\in[0,s_{\max}]}$ de durée $s_{\max}$ atteigne une région critique $\mathbf{F} \subset \mathcal{X}$ (que l'on suppose absorbante) de son espace d'état. Du point de vue applicatif, $X$ représente l'état du système industriel et l'événement $\{X \in \mathbf{F}\}$ correspond à la défaillance critique du système. On peut montrer (voir [8]) que la distribution d'importance optimale dans ce cas de figure se déduit directement de la fonction dite "committor" du processus :

$$\xi^*(x,s) = \mathbb{P}_{\mathbf{p}}\left(\exists t \in [0, s_{\max} - s] \,:\, X_{s+t} \in \mathbf{F} \mid X_s = x\right) \tag{3}$$

Il s'agit simplement de la probabilité que la trajectoire réalise l'événement d'intérêt avant la date $s_{\max}$ sachant son état $x$ à un instant donné $s$.

# 4 Approximation de la fonction committor

Cette fonction committor n'étant bien entendu pas connue, nous proposons de la remplacer par une approximation bien choisie. La fonction committor quantifie en un sens la proximité d'un état de l'espace $\mathcal{X}$ à la région $\mathbf{F}$. Une bonne approximation doit au moins pouvoir quantifier la proximité de chaque régime à la région $\mathcal{V}_{\mathbf{F}}$ définie par l'ensemble des régimes permettant d'accéder à $\mathbf{F}$.

On peut remarquer que l'évolution du régime d'un PDMP est une marche aléatoire, non nécessairement Markovienne, sur un graphe. On sait cependant (c'est une conséquence de la formule de Dynkin), que les temps moyens d'atteinte de n'importe quelle région d'un graphe peuvent être calculés explicitement pour une marche aléatoire homogène en résolvant un système linéaire. On se donne donc, avant toute simulation, une marche aléatoire Markovienne et homogène sur le graphe dont les sommets sont les régimes du PDMP, de préférence proche de la marche aléatoire non Markovienne décrite au-dessus. On détermine depuis chaque sommet $v \in \mathcal{V}$ le temps moyen d'atteinte $\rho_v$ de la région $\mathcal{V}_{\mathbf{F}}$ pour cette marche aléatoire homogène.

On construit alors la famille d'approximations $(\xi_{\boldsymbol{\theta}})_{\theta\in\Theta}$ de la fonction committor $\xi^*$, indexée par un paramètre $\boldsymbol{\theta} \in \Theta$ de dimension arbitraire $d_\Theta$, à partir des temps moyens d'atteinte $(\rho_v)_{v\in\mathcal{V}}$ selon la formule suivante :

$$\xi_{\boldsymbol{\theta}}(v) = \exp\left(-\sum_{i=1}^{d_\Theta} \theta_i \times \rho_v^i\right). \tag{4}$$

À chaque valeur de $\boldsymbol{\theta}$ correspond une approximation $\xi_{\boldsymbol{\theta}}$, à laquelle on peut associer une distribution d'importance possible $\mathbf{g}_{\boldsymbol{\theta}}$. La meilleure distribution d'importance au sein de la famille $(\mathbf{g}_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ est déterminée séquentiellement par minimisation d'entropie croisée. Notre méthode permet une réduction de variance d'un facteur supérieur à 10 000 par rapport à une méthode de Monte Carlo standard sur un cas d'essai industriel complexe.

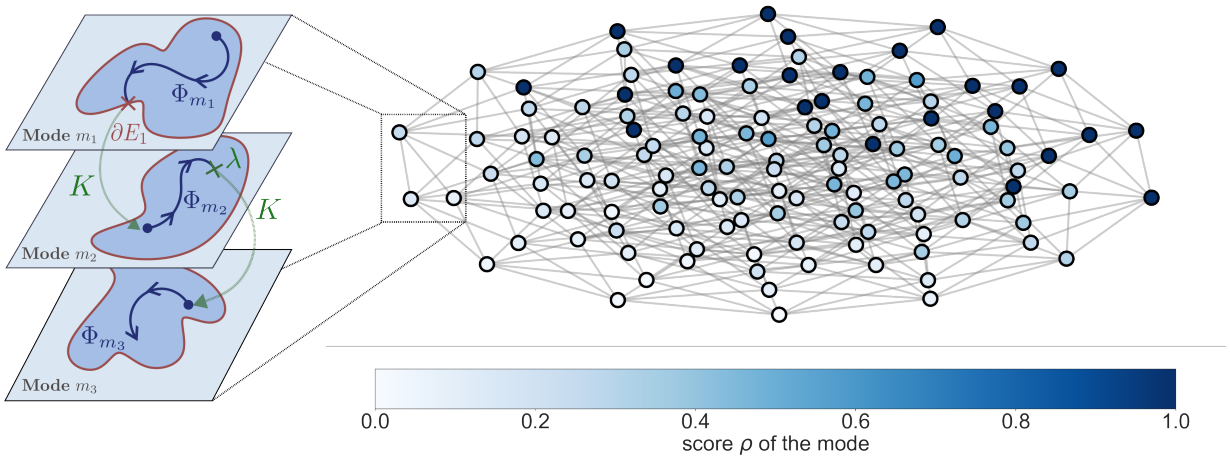



Figure 1: PDMP as a random walk on a graph

# References


[1] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

[2] Maxime El Masri. *High dimensional importance sampling through projections on a low dimensional subspace.* PhD thesis, ISAE-Institut Supérieur de l'Aéronautique et de l'Espace, 2022.

[3] Felipe Uribe, Iason Papaioannou, Youssef M Marzouk, and Daniel Straub. Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):818–847, 2021.

[4] Julien Demange-Chryst, François Bachoc, Jérôme Morio, and Timothé Krauth. Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling. *arXiv preprint arXiv:2310.09194*, 2023.

[5] Guillaume Chennetier, Hassane Chraibi, Anne Dutfoy, and Josselin Garnier. Adaptive importance sampling based on fault tree analysis for piecewise deterministic markov process. *arXiv preprint arXiv:2210.16185*, 2022.

[6] Mark HA Davis. Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):353–376, 1984.

[7] Benoite De Saporta, François Dufour, and Huilong Zhang. *Numerical methods for simulation and optimization of piecewise deterministic Markov processes: application to reliability.* John Wiley & Sons, 2015.

[8] Hassane Chraibi, Anne Dutfoy, Thomas Galtier, and Josselin Garnier. On the optimal importance process for piecewise deterministic markov process. *ESAIM: Probability and Statistics*, 23:893–921, 2019.

# Apprentissage en ligne

# Algorithmes de Newton stochastiques avec $O(Nd)$ opérations

Antoine Godichon-Baggioni [1] & Nicklas Werge [2]

[1] *LPSM, Sorbonne Université, France, antoine.godichon_baggioni@upmc.fr*
[2] *Department of Mathematics and Computer Science, University of Southern Denmark, werge@sdu.dk*

**Résumé.** On s'intéresse ici au traitement de données arrivant par blocs (en streaming)à l'aide d'algorithmes stochastiques dits adaptatifs. Plus précisément, on s'intéressera à des méthodes de Newton stochastiques, qui sont très utiles pour traiter des problèmes mal conditionnés. De plus, on verra que l'on peut obtenir de telles méthodes avec un temps de calculs de l'ordre de $O(Nd)$ opérations, i.e du même ordre que les algorithmes de gradient stochastiques classiques.

**Mots-clés.** Optimisation stochastique, méthodes adaptatives, algorithme de Newton, apprentissage en ligne.

**Abstract.** We focus on the processing of data arriving in blocks (streaming) using adaptive stochastic algorithm methods. Specifically, the focus is on stochastic Newton methods, which are highly useful for handling ill-conditioned problems. Furthermore, it will be shown that such methods can be achieved with a computational time of order $O(Nd)$ operations, i.e., of the same order as usual stochastic gradient algorithms.

**Keywords.** Stochastic optimization, adaptive methods, Newton's method, online learning.

## 1 Introduction

Un problème usuel consiste à estimer le minimiseur d'une fonction convexe $F : \mathbb{R}^d \to \mathbb{R}$ de la forme

$$F(\theta) := \mathbb{E}_{\xi\sim\Xi}[f(\theta;\xi)], \tag{1}$$

où $f$ est une fonction de perte, $\xi$ est une variable aléatoire suivant une distribution inconnue $\Xi$. Ce problème est fréquemment rencontré dans de nombreuses applications en apprentissage automatique [Kushner and Yin, 2003, Bottou et al., 2018].

Nous nous concentrons ici sur l'acquisition de données volumineuses arrivant en streaming. Plus précisément, les données arrivent sous forme de blocs [Godichon-Baggioni et al., 2023b, Godichon-Baggioni et al., 2023a], formant des sous-échantillons indépendants. Formellement, nous considérons une suite de copies i.i.d. : $\{\xi_{1,1},\dots,\xi_{1,n_1}\},\dots,\{\xi_{t,1},\dots,\xi_{t,n_t}\},\dots$, où $\{\xi_{t,1},\dots,\xi_{t,n_t}\}$ représente un bloc de $n_t$ données arrivant au temps $t$.

Nous nous intéressons ici à des méthodes de gradient stochastiques adaptatives, c'est-à-dire que nous incorporons une matrice aléatoire $A_t$ dans le pas. En particulier, nous étudierons le cas où $A_t$ est un estimateur de l'inverse de la Hessienne de $F$, correspondant à un algorithme de Newton en streaming. Ces méthodes sont particulièrement utiles lorsqu'il faut traiter des fonctions mal conditionnées, c'est-à-dire lorsque les valeurs propres de la Hessienne de la fonction à minimiser sont à des échelles très différentes [Bercu et al., 2020, Boyer and Godichon-Baggioni, 2023]. Ces méthodes adaptatives peuvent s'écrire de manière récursive comme suit :

$$\theta_{t+1} = \theta_t - \gamma_{t+1} A_t \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{2}$$

où $\nabla_\theta f(\theta_t; \xi_{t+1}) = n_{t+1}^{-1} \sum_{i=1}^{n_{t+1}} \nabla_\theta f(\theta_t, \xi_{t+1,i})$ et $(\gamma_t)$ est une suite de pas positifs.

Nous introduirons également une version moyennée pondérée de ces algorithmes [Polyak and Juditsky, 19 Mokkadem and Pelletier, 2011, Boyer and Godichon-Baggioni, 2023]. Cela permet, via la moyennisation, d'obtenir des estimateurs asymptotiquement efficaces, tandis que les pondérations permettent de donner plus d'importance aux derniers estimateurs de gradient et ainsi réduire les éventuels problèmes d'initialisation. Enfin, nous montrons que l'approche en streaming permet d'obtenir (dans certains cas tels que les régressions linéaires, logistiques et softmax, ainsi que l'estimation de la médiane) des estimateurs de Newton stochastiques ne nécessitant que $O(Nd)$ opérations (où $N$ est la taille totale de l'échantillon), c'est-à-dire aussi peu coûteux que les algorithmes de premier ordre tels que les algorithmes de gradient stochastiques ou Adagrad. Ces estimateurs sont donc "optimaux" en termes de temps de calculs, tout en restant asymptotiquement efficaces.

# 2 Cadre

Dans cette section, on introduit les hypothèses nécessaires pour l'obtention de nos résultats théoriques. Ces hypothèses sont usuelles en optimisation stochastique, et en particulier pour les méthodes adaptatives [Leluc and Portier, 2023, Boyer and Godichon-Baggioni, 2023, Kushner and Yin, Duflo, 2013, Godichon-Baggioni and Tarrago, 2023].

**Hypothèse 1** *Pour presque tout $\xi$, la fonction $f(\cdot; \xi)$ est différentiable et il existe des constantes positives $C$ et $C'$ telles que pour tout $\theta \in \mathbb{R}^d$*

$$\mathbb{E}[\|\nabla_\theta f(\theta; \xi)\|^2] \leq C + C'(F(\theta) - F(\theta^*)). \tag{3}$$

*De plus, il existe $\theta^* \in \mathbb{R}^d$ tel que $\nabla_\theta F(\theta^*) = 0$, et la fonction $\Sigma : \theta \to \mathbb{E}[\nabla_\theta f(\theta; \xi) \nabla_\theta f(\theta; \xi)^\top]$ est continue en $\theta^*$.*

A noter que dans l'Hypothèse 1, $\mathbb{E}[\|\nabla_\theta f(\theta; \xi)\|^2]$ n'est pas majoré par une constante auquel on ajoute l'erreur quadratique $\|\theta - \theta^*\|^2$. Au lieu de cela, nous utilisons l'erreur fonctionnelle $F(\theta) - F(\theta^*)$ [Gower et al., 2019, Gazagnadou et al., 2019]. Cependant, si la fonctionnelle $F$ est $\mu$ fortement convexe et $L_{\nabla F}$-lisse, on a $\frac{2}{L_{\nabla F}}(F(\theta) - F(\theta^*)) \leq \|\theta - \theta^*\|^2 \leq \frac{2}{\mu}(F(\theta) - F(\theta^*))$ pour tout $\theta \in \mathbb{R}^d$.

Afin d'assurer la forte consistance des estimateurs, nous introduisons une deuxième hypothèse. Cette dernière permet l'utilisation d'un développement de Taylor à l'ordre 2 de la fonctionnelle $F$.

**Hypothèse 2** *La fonctionnelle $F$ est deux fois continûment différentiable avec une hessienne uniformément bornée, i.e il existe $L_{\nabla F}$ tel que $\|\nabla^2_\theta F(\theta)\|_{\text{op}} \leq L_{\nabla F}$.*

A noter que cela implique, entre autre, que le gradient de $F$ est $L_{\nabla F}$-Lipschitz. La troisième hypothèse permet d'assurer l'unicité du minimiseur $\theta^*$ de la fonctionnelle $F$.

**Hypothèse 3** *La fonctionnelle $F$ est localement fortement convexe: $\lambda_{\min} := \lambda_{\min}(\nabla^2_\theta F(\theta^*)) > 0$.*

# 3 Méthodes adaptatives en streaming

Pour simplifier les résultats et notations, on considère maintenant que la taille des sous échantillons est constante, i.e $n_t = n$ pour tout $t \geq 0$. Néanmoins, les résultats pour des tailles de sous-échantillons croissantes sont disponibles dans [Godichon-Baggioni and Werge, 2023]. Dans tout ce qui suit, on note $N_t$ le nombre total de données traitées au temps $t$, i.e $N_t = nt$. On suppose également que $A_t$ est symétrique et définie positive pour tout $t \geq 0$. De plus, on suppose à partir de maintenant que la suite de pas $(\gamma_t)$ et la suite de matrices aléatoires $(A_t)$ vérifient les conditions suivantes:

$$\sum_{t\geq 1} \gamma_t \lambda_{\min}(A_{t-1}) = +\infty \text{ p.s.}, \quad \text{et} \quad \sum_{t\geq 1} \gamma_t^2 \lambda_{\max}(A_{t-1})^2 < +\infty \text{ p.s.} \tag{4}$$

Enfin, il existe une filtration $\mathcal{F}_t$ telle que $A_t$ soit $\mathcal{F}_t$ mesurable et $\xi_{t+1}$ soit indépendant de $\mathcal{F}_t$. Ces hypothèses sont assez usuelles et sont même vitales pour prouver la forte consistance des estimateurs à l'aide du théorème de Robbins-Siegmund [Boyer and Godichon-Baggioni, 2023]. Dans ce qui suit, on prendra $\gamma_t = C_\gamma t^{-\gamma}$ avec $C_\gamma > 0$ et $\gamma \in (1/2, 1)$.

Le théorème suivant établit la forte consistance de nos estimateurs de gradient stochastiques adaptatifs $(\theta_t)$ définis par (2).

**Théorème 1** *Supposons que les Hypothèses 1 à 3 sont vérifiées, ainsi que les conditions* (4)*. Alors, $\theta_t$ converge presque sûrement vers $\theta^*$.*

L'hypothèse suivante permet d'obtenir les vitesses de convergence des estimateurs $(\theta_t)$.

**Hypothèse 4** *La matrice aléatoire $A_t$ converge presque sûrement vers une matrice définie positive $A$.*

Par exemple, dans les méthodes de Newton, la matrice $A$ correspond à l'inverse de la Hessienne, et dans le cas d'Adagrad, elle correspond à l'inverse de la racine carrée de la diagonale de la variance du gradient. A noter également que la forte consistance de $\theta_t$ (cf Théorème 1) implique souvent la forte consistance de $A_t$.

**Théorème 2** *On suppose que les Hypothèses 1 à 4 sont vérifiées, ainsi que les conditions* (4)*. De plus, on suppose qu'il existe des constantes positives* $C_\eta$ *et* $\eta > \frac{1}{\gamma} - 1$ *telles que pour tout* $\theta \in \mathbb{R}^d$*,*

$$\mathbb{E}\left[\|\nabla_\theta f(\theta;\xi)\|^{2+2\eta}\right] \leq C_\eta \left(1 + F(\theta) - F(\theta^*)\right)^{1+\eta}. \tag{5}$$

*Alors,*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^\gamma}\right) \ p.s.$$

# 4 La version moyennée pondérée

La version moyennée pondérée des méthodes de gradient stochastiques adaptatives pour les données en streaming est définie comme suit:

$$\theta_{t,w} = \frac{1}{\sum_{i=0}^{t-1} \ln(i+1)^w} \sum_{i=0}^{t-1} \ln(i+1)^w \theta_i, \tag{6}$$

ce qui peut être écrit récursivement comme

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^{t} \ln(i+1)^w}(\theta_t - \theta_{t,w}).$$

Cette moyenne pondérée dans (6) améliore le comportement des estimateurs en attribuant plus de poids aux dernières estimations de $(\theta_t)$. La pondération logarithmique met donc l'accent sur les estimations récentes, présumées plus précises, tout en assurant l'efficacité des estimateurs [Mokkadem and Pelletier, 2011, Boyer and Godichon-Baggioni, 2023]. Pour établir la vitesse de convergence des estimateurs moyennés, on introduit une nouvelle hypothèse.

**Hypothèse 5** *Il existe des constantes positives* $L_r$ *et* $r$ *telles que pour tout* $\theta \in \mathcal{B}(\theta^*, r)$

$$\|\nabla_\theta F(\theta) - \nabla_\theta^2 F(\theta^*)(\theta - \theta^*)\| \leq L_r \|\theta - \theta^*\|^2.$$

Cette hypothèse est satisfaite dès que la Hessienne de $F$ est localement Lipschitzienne sur un voisinage de $\theta^*$.

**Théorème 3** *On suppose que les Hypothèses 1 à 5 sont vérifiées ainsi que l'inégalité* (5)*. De plus, on suppose qu'il existe* $v' > 1/2$ *tel que*

$$\frac{1}{\sum_{i=0}^{t-1} \ln(i+1)^w} \sum_{i=0}^{t-1} \ln(i+1)^{w+1/2+\delta} \|A_{i+1}^{-1} - A_i^{-1}\|_{\text{op}} (i+1)^{\frac{\gamma}{2}} = \mathcal{O}\left(\frac{1}{t^{v'}}\right) \ p.s, \tag{7}$$

*pour n'importe quel* $\delta > 0$*. Alors*

$$\|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \ p.s \quad et \quad \sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}).$$

La variance $\nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}$ correspond à l'inverse de l'Information de Fisher, et les estimateurs sont donc asymptotiquement efficaces.

# 5 Applications aux algorithmes de Newton

## 5.1 Algorithmes de Newton en streaming

Dans le cas particulier des méthodes de Newton stochastiques, on peut obtenir l'efficacité asymptotique sans moyennisation en choisissant une suite de pas de la forme $\gamma_t = \frac{1}{t}$. L'algorithme de Newton stochastique est alors défini de manière récursive pour tout $t \geq 0$ par

$$\theta_{t+1} = \theta_t - \frac{1}{t+1}\overline{H}_t^{-1}\nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{8}$$

où $\nabla_\theta f(\theta_t; \xi_{t+1}) = n^{-1}\sum_{i=1}^n \nabla_\theta f(\theta_t; \xi_{t+1,i})$ et $\overline{H}_t$ est un estimateur de la hessienne de $F$. De plus, on supposera que $\overline{H}_t$ est de la forme $\overline{H}_t = N_t^{-1}H_t$ avec

$$H_t = H_0 + \sum_{i=1}^{t}\sum_{j=1}^{n}\alpha_{i,j}\Phi_{i,j}\Phi_{i,j}^\top,$$

avec $H_0$ symétrique et positive, $\alpha_{i,j} = \alpha(\theta_{i-1}; \xi_{i,j})$, et $\Phi_{i,j} = \Phi(\theta_{i-1}; \xi_{i,j})$. On peut alors mettre à jour $H_t^{-1}$ de manière récursive et avec un temps de calculs réduit à l'aide de la formule de Riccati/Sherman-Morrisson [Duflo, 2013, Sherman and Morrison, 1950] utilisée $n$ fois, i.e pour tout $j = 1, \ldots, n$,

$$H_{t-1,j}^{-1} = H_{t-1,j-1}^{-1} - \alpha_{t,j}\left(1 + \alpha_{t,j}\Phi_{t,j}^\top H_{t-1,j-1}^{-1}\Phi_{t,j}\right)^{-1} H_{t-1,j-1}^{-1}\Phi_{t,j}\Phi_{t,j}^\top H_{t-1,j-1}^{-1}.$$

avec la convention $H_{t-1,0}^{-1} = H_{t-1}$. La construction explicite des estimations récursives de l'inverse de la hessienne est détaillée dans diverses applications, notamment les régressions linéaires, logistiques, softmax et ridge [Bercu et al., 2020, Boyer and Godichon-Baggioni, 2023, Godichon-Baggioni et al., 2024].

**Théorème 4** *On suppose que les Hypothèses 1, 2, 3 et 5 sont vérifiées, ainsi que l'inégalité* (5)*. Alors, $\theta_t$ converge presque sûrement vers $\theta^*$. De plus, supposons que $\overline{H}_t^{-1}$ converge presque sûrement vers $\nabla_\theta^2 F(\theta^*)^{-1}$, alors*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \ p.s.$$

*Suppose de plus qu'il existe une constante positive $\nu$ telle que $\|\overline{H}_t^{-1} - \nabla_\theta^2 F(\theta^*)^{-1}\|_{op} = \mathcal{O}(\frac{1}{t^\nu})$ p.s. Alors*

$$\sqrt{N_t}(\theta_t - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}).$$

## 5.2 Algorithmes de Newton stochastiques avec $\mathcal{O}(dN_t)$ opérations

La méthode de Newton stochastique nécessite $\mathcal{O}(d^2 N_t)$ opérations, ce qui peut être coûteux en terme de temps de calcul, surtout dans des contextes de (relativement) grande dimension.

Pour pallier ce problème, on remplace $H_t$ dans (8) par $H_{t,w'}$ défini par

$$H_{t,w'} = H_{0,w'} + \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} Z_{i,j} \left( \iota_{i,j} \tilde{e}_{i,j} \tilde{e}_{i,j}^\top + \alpha_{i,j} \Phi_{i,j} \Phi_{i,j}^\top \right), \tag{9}$$

avec $N_{t,Z} = 1 + \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} Z_{i,j}$, $H_{0,w'}$ symétrique et positive, $w' \geq 0$, et $Z_{i,j}$ i.i.d avec $Z_{i,j} \sim \mathcal{B}(p)$ pour un certain $p \in (0,1]$. De plus, soit $N_{t,k,Z} = (1 + \sum_{i=1}^{t-1} \sum_{j=1}^{n} Z_{i,j} + \sum_{j=1}^{k} Z_{t,j})$, $\iota_{i,j} = c_\iota N_{i,j,Z}^{-\iota}$ pour $\iota \in (0, 1/2)$, et $e_{i,j}$ soit le composant ($N_{i,j,Z}$ modulo $d+1$)-ème de la base canonique. A noter

$$\tilde{H}_{t,j,w'}^{-1} = H_{t,j-1,w'}^{-1} - \frac{Z_{t+1,j} \iota_{t+1,j}}{1 + \iota_{t+1,j} e_{t+1,j} \tilde{H}_{t,j-1,w'}^{-1} e_{t+1,j}^T} H_{t,j-1,w'}^{-1} e_{t+1,j} e_{t+1,j}^T H_{t,j,w'}^{-1}$$

$$H_{t,j,w'}^{-1} = H_{t,j-1,w'}^{-1} - \frac{Z_{t+1,j} \ln(t+1)^{w'} \alpha_{t+1,j}}{1 + \ln(t+1)^{w'} \alpha_{t+1,j} \Phi_{t+1,j}^T H_{t,j-1,w'}^{-1} \Phi_{t+1,j}} H_{t,j-1,w'}^{-1} \Phi_{t+1,j} \Phi_{t+1,j}^T H_{t,j-1,w'}^{-1}$$

avec $\tilde{H}_{t,0,w'}^{-1} = H_{t-1,w'}^{-1}$ et $H_{t,0,w'}^{-1} = \tilde{H}_{t,n,w'}^{-1}$. La mise à jour de $H_{t+1}^{-1}$ ne coûte en moyenne que $\mathcal{O}(pd^2n)$ opérations, conduisant à un nombre total d'opérations d'ordre (en moyenne)

$$\underbrace{pd^2 N_t}_{\text{estimation de l'inverse de la Hessienne}} + \underbrace{dN_t}_{\text{estimation du gradient}} + \underbrace{\frac{d^2 N_t}{n}}_{\text{multiplication Hessienne*gradient}} .$$

Ainsi, on peut jouer avec la valeur de $p$ pour réduire le coût de la mise à jour de l'inverse de la Hessienne. En effet, on peut obtenir un coût computationnel moyen au temps $t$ de l'ordre de $\mathcal{O}(dN_t)$ opérations en prenant $p = d^{-1}$ et $n = d$. En d'autres termes, il est possible d'obtenir une méthode de Newton stochastique avec seulement $\mathcal{O}(dN_t)$ opérations.

Dans ce qui suit, on suppose que pour tout $\theta \in \mathbb{R}^d$,

$$\nabla_\theta^2 F(\theta) = \mathbb{E}\left[ \alpha(\theta;\xi) \Phi(\theta;\xi) \Phi(\theta;\xi)^T \right]. \tag{10}$$

**Théorème 5** *On suppose que les Hypothèses 1, 2, 3 et 5 sont vérifiées, ainsi que les inégalités* (5) *et* (10)*. De plus, supposons que pour tout $\theta$, il existe des constantes positives $C_{\eta'}$ et $\eta' > 1$ telles que pour tout $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'}] \leq C_{\eta'}^{\eta'}.$$

*Alors,*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left( \frac{\ln(N_t)}{N_t} \right) \ p.s.$$

*De plus, on suppose que la Hessienne de $F$ est localement $L_{\nabla^2 F}$-Lipschitz sur un voisinage autour de $\theta^*$ et que $\eta' \geq 2$. Alors*

$$\sqrt{N_t}(\theta_t - \theta^*) \xrightarrow[t \to +\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}).$$

### 5.3 Version moyennée pondérée de la méthode de Newton stochastique avec $\mathcal{O}(dN_t)$ opérations

Bien que la méthode de Newton stochastique "directe" soit très performante, elle peut être assez sensible à une mauvaise initialisation [Boyer and Godichon-Baggioni, 2023]. On considère donc une version moyennée pondérée définie récursivement par

$$\theta_{t+1} = \theta_t - \gamma_{t+1}\overline{S}_{t,w'}^{-1}\nabla_\theta f(\theta_t;\xi_{t+1}),\ \theta_0 \in \mathbb{R}^d, \tag{11}$$

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^t \ln(i+1)^w}(\theta_t - \theta_{t,w}), \tag{12}$$

$\overline{S}_{t,w'} = N_{t,Z}^{-1}S_{t,w'}$ avec $N_{t,Z}\overline{S}_{t,w'} =: S_{t,w'} = S_{0,w'} + \sum_{i=1}^t \ln(i+1)^{w'}\sum_{j=1}^n Z_{i,j}\left(\iota_i e_{i,j}e_{i,j}^\top + \alpha_{i,j}\Phi_{i,j}\Phi_{i,j}^\top\right)$ et $S_{0w'}$ symétrique et positive. On peut suivre le même schéma récursif que pour $H_{t,w'}^{-1}$ pour mettre à jour l'inverse de $S_{t,w'}^{-1}$. En effet, la seule différence entre $S_{t,w'}^{-1}$ et $H_{t,w'}^{-1}$ est le choix de l'estimateur de $\theta^*$ choisi pour la méthode du plug-in.

**Théorème 6** *On suppose que les hypothèses 1, 2, 3 et 5 sont vérifiées, ainsi que les inégalités* (5) *and* (10)*. De plus, supposons que pour tout $\theta$, il existe des constantes positives $C_{\eta'}$ et $\eta' > 1$ telles que pour tout $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'}] \leq C_{\eta'}^{\eta'}.$$

*Alors,*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^\gamma}\right)\ p.s., \qquad \|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right)\ p.s.,$$

*et*

$$\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}).$$

# 6 Simulations

Dans cette section, on se concentre sur deux exemples classiques: la régression linéaire et la régression logistique. Pour le modèle linéaire, on a $\xi = (x, y) \in \mathbb{R}^d \times \mathbb{R}$, et on cherche à minimiser la fonction $F(\theta) = \frac{1}{2}\mathbb{E}[(y - x^\top\theta)^2]$. Dans le cas de la régression logistique, on a $\xi = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$, et la fonction correspondante est $F(\theta) = \mathbb{E}[\ln(1+\exp(x^\top\theta)) - yx^\top\theta]$. Dans les deux cas, on va considérer une structure de covariance "complexe", i.e on va prendre

$$x \sim \mathcal{N}\left(0, M\,\mathrm{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d} M^\top\right).$$

où $M$ est une matrice orthogonale. Ce choix de distribution nous permet d'introduire des corrélations fortes entre les coordonnées de $x$. Dans ce qui suit, on fixe $d = 100$ et on note donc que la Hessienne a des valeurs propres à des échelles très différentes.

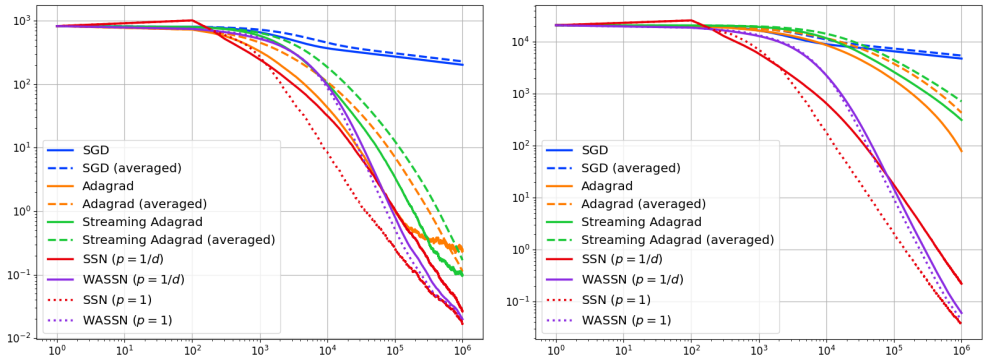



Figure 1: Modèle linéaire : Evolution de l'erreur quadratique moyenne des estimateurs en fonction de la taille d'échantillon. Les points initiaux $\theta_0$ sont définis par $\theta_0 = \theta^*(1 + rU)$, où $U$ suit une loi uniforme sur la sphère unitaire de $\mathbb{R}^d$, et $r = 1$ (à gauche) ou $r = 5$ (à droite).

### 6.1 Modèle linéaire

Dans la Figure 1, on considère le modèle linéaire avec deux types d'initialisations (plus ou moins précises). On peut voir que Adagrad et les algorithmes de Newton présentent des taux de convergence plus rapides par rapport au SGD standard (sans surprise). A noter que bien que l'algorithme Adagrad adapte ses pas, il peut être moins efficace lorsqu'il est confronté à des données fortement corrélées. C'est particulièrement criant lorsque les algorithmes sont mal initialisés, tandis les deux méthodes de Newton restent très performantes.

### 6.2 Régression logistique

Dans la Figure 2, on considère le problème de régression logistique avec là encore deux initialisations différentes. Pour toutes les configurations initiales, les méthodes de Newton stochastiques sont particulièrement efficace tandis qu'Adagrad semble moins adaptés.

## Bibliographie

## References

[Bercu et al., 2020] Bercu, B., Godichon, A., and Portier, B. (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367.

[Bottou et al., 2018] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

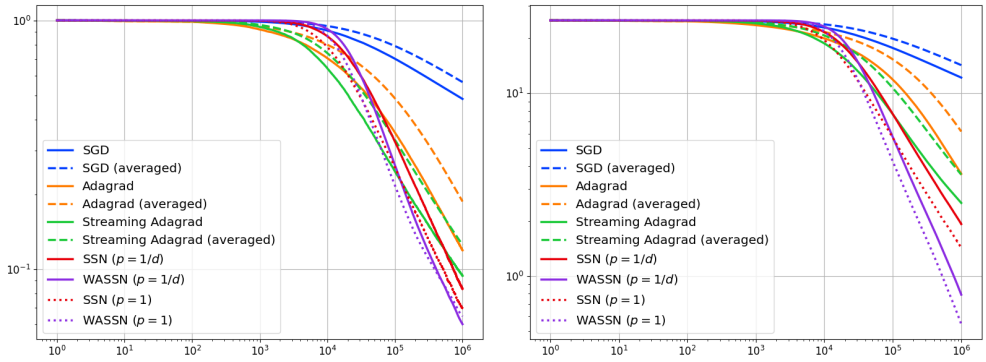



Figure 2: Régression logistique : Evolution de l'erreur quadratique moyenne des estimateurs en fonction de la taille d'échantillon. Les points initiaux $\theta_0$ sont définis par $\theta_0 = \theta^*(1 + rU)$, où $U$ suit une loi uniforme sur la sphère unitaire de $\mathbb{R}^d$, et $r = 1$ (à gauche) ou $r = 5$ (à droite).

[Boyer and Godichon-Baggioni, 2023] Boyer, C. and Godichon-Baggioni, A. (2023). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972.

[Duflo, 2013] Duflo, M. (2013). *Random iterative models*, volume 34. Springer Science & Business Media.

[Gazagnadou et al., 2019] Gazagnadou, N., Gower, R., and Salmon, J. (2019). Optimal mini-batch and step sizes for saga. In *International conference on machine learning*, pages 2142–2150. PMLR.

[Godichon-Baggioni et al., 2024] Godichon-Baggioni, A., Lu, W., and Portier, B. (2024). Recursive ridge regression using second-order stochastic algorithms. *Computational Statistics & Data Analysis*, 190:107854.

[Godichon-Baggioni and Tarrago, 2023] Godichon-Baggioni, A. and Tarrago, P. (2023). Non asymptotic analysis of adaptive stochastic gradient algorithms and applications. *arXiv preprint arXiv:2303.01370*.

[Godichon-Baggioni and Werge, 2023] Godichon-Baggioni, A. and Werge, N. (2023). On adaptive stochastic optimization for streaming data: A newton's method with o (dn) operations. *arXiv preprint arXiv:2311.17753*.

[Godichon-Baggioni et al., 2023a] Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2023a). Learning from time-dependent streaming data with online stochastic algorithms. *Transactions on Machine Learning Research*.

[Godichon-Baggioni et al., 2023b] Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2023b). Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *ESAIM: Probability and Statistics*, 27:482–514.

[Gower et al., 2019] Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR.

[Gower et al., 2021] Gower, R. M., Richtárik, P., and Bach, F. (2021). Stochastic quasi-gradient methods: Variance reduction via jacobian sketching. *Mathematical Programming*, 188:135–192.

[Kushner and Yin, 2003] Kushner, H. and Yin, G. (2003). *Stochastic approximation and recursive algorithms*. Springer-Verlag NY.

[Leluc and Portier, 2023] Leluc, R. and Portier, F. (2023). Asymptotic analysis of conditioned stochastic gradient descent. *Transactions on Machine Learning Research*.

[Mokkadem and Pelletier, 2011] Mokkadem, A. and Pelletier, M. (2011). A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4):1523–1543.

[Polyak and Juditsky, 1992] Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30:838–855.

[Sherman and Morrison, 1950] Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.

# Estimation en ligne de l'inverse de la Hessienne pour l'optimisation stochastique avec application aux algorithmes de Newton stochastiques universels

Antoine Godichon-Baggioni [1] , Wei Lu [2] , Bruno Portier [2]

[1] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, Paris, France.*
*antoine.godichon_baggioni@upmc.fr*
[2] *Laboratoire de Mathématiques de l'INSA Rouen Normandie (LMI), Normandie Université, Rouen, France.*
*wei.lu@insa-rouen.fr ; bruno.portier@insa-rouen.fr*

**Résumé.** On propose un algorithme stochastique du second-ordre (de type Newton) pour estimer le minimiseur d'une fonction convexe écrite comme une espérance. Nous introduisons une technique d'estimation récursive directe pour la matrice inverse de la Hessienne en utilisant une procédure de Robbins-Monro. Cette approche permet de réduire la complexité computationnelle. Surtout, elle permet de développer des méthodes de Newton stochastiques universelles tout en assurant l'efficacité asymptotique des estimateurs obtenus.

**Mots-clés.** Algorithme de Newton stochastique; Optimisation Stochastique; Algorithme de Robbins-Monro; Estimation en ligne

**Abstract.** This work addresses second-order stochastic optimization for estimating the minimizer of a convex function written as an expectation. A direct recursive estimation technique for the inverse Hessian matrix using a Robbins-Monro procedure is introduced. This approach enables to drastically reduce computational complexity. Above all, it allows to develop universal stochastic Newton methods and investigate the asymptotic efficiency of the proposed estimates.

**Keywords.** Stochastic Newton algorithm; Stochastic Optimization; Robbins-Monro algorithm; Online estimation

# 1 Introduction

Dans ce travail, nous considérons le problème d'optimisation stochastique, consistant à estimer le paramètre $\theta \in \mathbb{R}^d$ défini par

$$\theta = \arg\min_{h \in \mathbb{R}^d} G(h).$$

La fonction $G$ est définie pour tout $h \in \mathbb{R}^d$ par : $G(h) = \mathbb{E}\left[g(X, h)\right]$ où $X$ est un vecteur aléatoire de $\mathbb{R}^p$ et $g : \mathbb{R}^p \times \mathbb{R}^d \longrightarrow \mathbb{R}$ est une fonction connue, supposée deux fois différentiable.

Dans ce qui suit, on notera $\nabla_h g$ et $\nabla_h^2 g$, le gradient et la matrice Hessienne de $g$ par rapport à la seconde variable $h$, et $\nabla G$ et $\nabla^2 G$, le gradient et la matrice Hessienne de $G$. On suppose que la matrice $\nabla^2 G(\theta)$ est définie positive.

Nous nous intéressons à l'estimation récursive (ou en ligne) du paramètre $\theta$ à partir d'une suite de vecteurs aléatoires indépendants $(X_n)_{n\geq 1}$ ayant la même distribution que $X$. L'une des méthodes les plus connues dans ce contexte est l'algorithme du gradient stochastique, défini de manière récursive pour tout $n \geq 1$ par :

$$\theta_n^{SG} = \theta_{n-1}^{SG} - \nu_n \nabla_h g\left(X_n, \theta_{n-1}^{SG}\right)$$

où $\theta_0^{SG}$ est une valeur initiale choisie arbitrairement et $(\nu_n)_{n\geq 1}$ est une suite de nombres réels positifs décroissant vers 0.

Malgré leur efficacité reconnue, ces méthodes peuvent être très sensibles aux problèmes dit "mal conditionné", où la Hessienne a des valeurs propres à différentes échelles (voir par exemple, Bercu et al., 2020; Leluc et Portier, 2023). Pour surmonter ce problème, des algorithmes stochastiques du second ordre de la forme

$$\theta_n = \theta_{n-1} - \nu_n \, A_n \nabla_h g(X_n, \theta_{n-1})$$

ont été proposés et récemment étudiés. Ici, $(\nu_n)_{n\geq 1}$ est une suite de nombres réels positifs décroissant vers 0 et la matrice $A_n$ est un estimateur récursif de l'inverse de la matrice Hessienne de $G$ en $\theta$, c'est-à-dire un estimateur récursif de $H^{-1}$ avec $H = \nabla^2 G(\theta)$. La principale difficulté réside donc dans la construction de cette suite $A_n$.

Plusieurs algorithmes récursifs du second ordre ont été proposés et étudiés. Par exemple, Bercu et al. (2020) proposent un algorithme de Newton stochastique efficace pour estimer les paramètres d'un modèle de régression logistique. Dans un travail récent, Bercu et al. (2023) proposent un algorithme de Gauss-Newton stochastique pour estimer le coût de Transport Optimal entre deux mesures de probabilité discrètes. Godichon-Baggioni et al. (2024) proposent des algorithmes du second ordre pour résoudre le problème de régression Ridge dans le cadre linéaire et logistique, tandis que le cas de la médiane géométrique est introduit et étudié par Godichon-Baggioni et Lu (2023). Dans tous ces algorithmes, les estimateurs de l'inverse de la matrice Hessienne sont mis à jour de manière récursive en utilisant la formule d'inversion de Riccati (également appelée formule de Sherman-Morrison, voir par exemple Duflo (1996)). Ce calcul est rendu possible grâce à la forme particulière de l'estimateur de la matrice Hessienne $H$, présentée comme $(1/n)\sum_{k=1}^{n} a_k \phi_k \phi_k^T$, où $(a_n)_{n\geq 1}$ est une suite de variables aléatoires réelles positives et $(\phi_n)_{n\geq 1}$ est une suite de vecteurs aléatoires dans $\mathbb{R}^d$.

Cependant, il n'est pas toujours possible d'obtenir un tel estimateur de la matrice Hessienne. Dans ce travail, nous proposons de construire un estimateur récursif de $H^{-1}$ sans tenter d'abord de construire un estimateur de $H$. Cette approche est basée sur le fait que nous avons $HH^{-1} = H^{-1}H = I_d$ et, par conséquent, la relation suivante :

$$\mathbb{E}\left[H^{-1}\nabla_h^2 g(X,\theta) + \nabla_h^2 g(X,\theta)H^{-1} - 2I_d\right] = 0 \tag{1}$$

où $I_d$ désigne la matrice identité d'ordre $d$. En utilisant un algorithme de type Robbins-Monro, nous proposons un estimateur récursif de la matrice $H^{-1}$ définie pour tout $n \geq 1$

par

$$A_n = A_{n-1} - \gamma_n \left( A_{n-1} \nabla_h^2 g(X_n, \theta_{n-1}) + \nabla_h^2 g(X_n, \theta_{n-1}) A_{n-1} - 2\, I_d \right)$$

où $(\gamma_n)_{n\geq 1}$ est une suite de nombres réels positifs décroissant vers 0 et $\theta_{n-1}$ est un estimateur de $\theta$.

Nous obtenons ainsi un estimateur universel de l'inverse de la Hessienne. Après avoir apporté de légères modifications pour réduire la complexité du calcul et pour contrôler les valeurs propres de $A_n$, nous établissons la vitesse de convergence presque sûre pour l'estimateur proposé. Ces résultats restent vrais pour tout estimateur consistant de $\theta_n$. Sur la base de ce concept, nous introduisons un algorithme de Newton stochastique universel. Pour améliorer davantage la vitesse de convergence, nous considérons également sa version moyennée pondérée, comme discuté par Mokkadem et Pelletier(2011); Boyer et Godichon-Baggioni (2023). Enfin, nous donnerons leurs vitesses de convergence et démontrons l'efficacité asymptotique des estimateurs moyennés pondérés.

# 2 Contexte

Nous considérons le problème de minimisation de la fonction convexe $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par :

$$G(h) := \mathbb{E}\left[ g(X, h) \right],$$

où la perte $g(X, \cdot)$ est une fonction convexe, dérivable deux fois et $X$ est un vecteur aléatoire de $\mathbb{R}^p$. Nous supposons qu'il existe une valeur unique $\theta \in \mathbb{R}^d$ telle que

$$\nabla G(\theta) = 0.$$

Introduisons maintenant les hypothèses pour le cadre d'estimation du paramètre $\theta$:

**(A1)** Il existe $C > 0$ tel que pour tout $h \in \mathbb{R}^d$,

$$\mathbb{E}\left[ \|\nabla_h g(X, h)\|^2 \right] \leq C \left( 1 + G(h) - G(\theta) \right).$$

**(A2)** La fonction $G$ est deux fois différentiable et $\nabla^2 G(\theta)$ est définie positive. De plus, la Hessienne est uniformément bornée, c'est-à-dire qu'il existe une constante positive $L_{\nabla G}$ telle que pour tout $h \in \mathbb{R}^d$,

$$\left\| \nabla^2 G(h) \right\|_{op} \leq L_{\nabla G}.$$

**(A3)** La fonction $\nabla^2 G$ est Lipschitzienne dans un voisinage de $\theta$, c'est-à-dire qu'il existe des constantes positives $r > 0$ et $L_r$ telles que pour tout $h \in \mathcal{B}(\theta, r)$

$$\left\| \nabla^2 G(h) - \nabla^2 G(\theta) \right\|_{op} \leq L_r \left\| \theta - h \right\|,$$

où $\mathcal{B}(\theta, r)$ désigne une boule de rayon $r$ centrée en $\theta$.

**(A4)** Il existe $q > 2$ et $C_q$ tel que pour tout $h \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\|\nabla_h^2 g(X,h)\right\|_F^q\right] \leq C_q.$$

Ces hypothèses sont très proches de celles présentées dans la littérature (Pelletier, 2000; Gadat et Panloup, 2017; Godichon-Baggioni, 2019).

# 3 Estimation de l'inverse de la Hessienne

On s'intéresse dans cette section à l'estimation de l'inverse de la Hessienne de la fonction $G$ en $\theta$, noté $H^{-1}$ où $H = \nabla^2 G(\theta)$. Soit $(X_n)_{n\geq 1}$ une suite de vecteurs aléatoires indépendants dans $\mathbb{R}^p$ ayant la même distribution que $X$. Supposons d'abord que $\theta$ est connu. À partir de l'égalité (1), la matrice $H^{-1}$ satisfait une équation de la forme $\Phi(H^{-1}) = 0$. Nous pouvons alors employer la procédure de Robbins-Monro (Robbins et Monro, 1951) pour estimer récursivement le paramètre $H^{-1}$. En désignant cet estimateur par $\hat{A}_n$, pour tout $n \geq 1$, nous avons :

$$\hat{A}_n = \hat{A}_{n-1} - \gamma_n \left(\hat{A}_{n-1}\nabla_h^2 g(X_n,\theta) + \nabla_h^2 g(X_n,\theta)\hat{A}_{n-1} - 2\, I_d\right),$$

où $\hat{A}_0$ est une matrice symétrique définie positive choisie arbitrairement, et $\gamma_n = c_\gamma n^{-\gamma}$ avec $\frac{1}{2} < \gamma < 1$ et $c_\gamma > 0$. Il est important de noter que par construction, la matrice $\hat{A}_n$ est symétrique pour tout $n \geq 1$. Cependant, puisque $\theta$ est inconnu, nous devons l'estimer. En supposant que nous disposons d'un estimateur récursif efficace de $\theta$ (par exemple, un estimateur du gradient stochastique), nous pouvons facilement déduire un estimateur de $H^{-1}$ en utilisant une procédure de substitution :

$$\hat{A}_n = \hat{A}_{n-1} - \gamma_n \left(\hat{A}_{n-1}\nabla_h^2 g(X_n,\tilde{\theta}_{n-1}) + \nabla_h^2 g(X_n,\tilde{\theta}_{n-1})\hat{A}_{n-1} - 2\, I_d\right).$$

Cet estimateur est toujours symétrique mais pas nécessairement défini positif. Pour garantir cette dernière propriété, nous introduisons une troncature basée sur la norme de $\nabla_h^2 g(X_n,\tilde{\theta}_{n-1})$, conduisant à l'estimateur suivant de $H^{-1}$ :

$$\hat{A}_n = \hat{A}_{n-1} - \gamma_n \left(\hat{A}_{n-1}\nabla_h^2 g(X_n,\tilde{\theta}_{n-1}) + \nabla_h^2 g(X_n,\tilde{\theta}_{n-1})\hat{A}_{n-1} - 2I_d\right) \mathbb{1}_{\{\|\nabla_h^2 g(X_n,\tilde{\theta}_{n-1})\|_{op}\leq \beta_n\}}, \tag{2}$$

où $\beta_n = c_\beta n^\beta$ avec $\frac{1-\gamma}{q-1} < \beta < \gamma - \frac{1}{2}$ et $0 < c_\gamma c_\beta < \frac{1}{2}$. En outre, cette troncature permet de contrôler la plus petite valeur propre de $\hat{A}_n$, ce qui est utile pour étudier un estimateur du paramètre $\theta$ impliquant la matrice $\hat{A}_n$. Cela est particulièrement important pour établir la consistance de l'algorithme de Newton stochastique présenté dans la Section 4.

Cependant, bien que cet estimateur soit efficace, chaque mise à jour nécessite des multiplications matricielles, induisant une complexité computationnelle de l'ordre de $\mathcal{O}(d^3)$, qui est la même que pour l'inversion matricielle. Néanmoins, nous pouvons introduire un algorithme d'une complexité en $\mathcal{O}(d^2)$, basé sur l'observation suivante : soit $Z$ un vecteur aléatoire centré dans $\mathbb{R}^d$ avec une matrice de variance-covariance $I_d$, indépendant du vecteur $X$. Alors,

$$\mathbb{E}\left[H^{-1}\, Z\, Z^T \nabla_h^2 g(X,\theta) \; + \nabla_h^2 g(X,\theta) Z\, Z^T H^{-1} \; - \; 2I_d\right] = 0.$$

Ainsi, en considérant une suite $(Z_n)_{n\geq 1}$ de vecteurs aléatoires indépendants et identiquement distribués bornés de $\mathbb{R}^d$ tels que $\mathbb{E}[Z_n] = 0$ et $\mathbb{E}\left[Z_n Z_n^T\right] = I_d$, et indépendants de $(X_n)_{n\geq 1}$, nous pouvons proposer un autre estimateur de $H^{-1}$ défini pour tout $n \geq 1$ comme suit :

$$\begin{aligned} P_n &= A_{n-1} Z_n \\ Q_n &= \nabla_h^2 g(X_n, \tilde{\theta}_{n-1}) Z_n \\ A_n &= A_{n-1} - \gamma_n \left( P_n Q_n^T + Q_n P_n^T - 2\, I_d \right) \mathbb{1}_{\{M\|Q_n\| \leq \beta_n\}} \end{aligned} \tag{3}$$

où $A_0$ est une matrice symétrique et définie positive et $M$ vérifie $\|Z_n\| \leq M$.

# 4 Algorithme de Newton Stochastique Universel

Dans cette section, nous introduisons l'Algorithme de Newton Stochastique Universel défini pour tout $n \geq 1$ par

$$\begin{aligned} \widehat{P}_n &= \widehat{A}_{n-1} Z_n \\ \widehat{Q}_n &= \nabla_h^2 g(X_n, \widehat{\theta}_{n-1}) Z_n \\ \widehat{A}_n &= \widehat{A}_{n-1} - \gamma_n \left( \widehat{P}_n \widehat{Q}_n^T + \widehat{Q}_n \widehat{P}_n^T - 2 I_d \right) \mathbb{1}_{M\|Q_n\| \leq \beta_n} \\ \widehat{\theta}_n &= \widehat{\theta}_{n-1} - \nu_n \widehat{A}_{n-1} \nabla_h g(X_n, \widehat{\theta}_{n-1}). \end{aligned}$$

En suivant le même schéma de preuve que pour le Théorème 4.2, on peut montrer que :

$$\|\widehat{\theta}_n - \theta\|^2 = O\left( \frac{\ln n}{n^\nu} \right) \quad p.s.$$

De plus, suivant Mokkadem et Pelletier (2011); Boyer et Godichon-Baggioni (2022), nous pouvons proposer un algorithme de Newton Stochastique Universel Moyenné Pondéré. Il est donné par

$$\begin{aligned} P_n &= A_{n-1} Z_n \\ Q_n &= \nabla_h^2 g(X_n, \theta_{n-1,\tau'}) Z_n \\ \theta_n &= \theta_{n-1} - \nu_n A_{n-1,\tau} \nabla_h g(X_n, \theta_{n-1}) && (4) \\ \theta_{n,\tau'} &= \left( 1 - \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^n \ln(k+1)^{\tau'}} \right) \theta_{n-1,\tau'} + \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^n \ln(k+1)^{\tau'}} \theta_n && (5) \\ A_n &= A_{n-1} - \gamma_n \left( P_n Q_n^T + Q_n P_n^T - 2\, I_d \right) \mathbb{1}_{M\|Q_n\| \leq \beta_n} && (6) \\ A_{n,\tau} &= (1 - \frac{\ln(n+1)^{\tau}}{\sum_{k=0}^n \ln(k+1)^{\tau}}) A_{n-1,\tau} + \frac{\ln(n+1)^{\tau}}{\sum_{k=0}^n \ln(k+1)^{\tau}} A_n && (7) \end{aligned}$$

où $(\nu_n)_{n\geq 1}$ est une suite de nombres réels positifs définie pour tout $n \geq 1$ par $\nu_n = c_\nu n^\nu$ avec $c_\nu > 0$ et $\nu \in (1/2, 1-\beta)$ satisfaisant $\gamma + \nu > 3/2$. De plus, $\tau, \tau' \geq 0$. Le théorème suivant donne la consistance des estimateurs définis par (4) et (5).

**Theorem 4.1** *Supposons que les hypothèses* ***(A1)*** *à* ***(A4)*** *soient vérifiées. Soient* $\theta_n$ *et* $\theta_{n,\tau'}$ *définis comme dans* (4) *et* (5)*. Alors,*

$$\theta_n \xrightarrow[n\to\infty]{p.s.} \theta \quad et \quad \theta_{n,\tau'} \xrightarrow[n\to\infty]{p.s.} \theta.$$

Notez que la contrainte $\gamma + \nu > 3/2$ est de nature purement technique mais est cruciale pour l'application du Théorème de Robbins-Siegmund et donc pour obtenir la consistance des estimateurs. Cependant, nous pensons que cette condition pourrait ne pas être nécessaire en pratique. Nous pouvons maintenant donner les vitesses de convergence presque sûr des estimateurs.

**Theorem 4.2** *Supposons que les hypothèses* ***(A1)*** *à* ***(A4)*** *soient vérifiées. Alors* $\theta_n$ *et* $\theta_{n,\tau'}$ *définis par* (4) *et* (5) *vérifient pour tout* $\delta > 0$

$$\|\theta_n - \theta\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\nu}}\right) p.s. \quad et \quad \|\theta_{n,\tau'} - \theta\|^2 = o\left(\frac{\ln n^{1+\delta}}{n}\right) p.s.$$

*De plus,* $A_n$ *et* $A_{n,\tau}$ *définis par* (6) *et* (7) *vérifient*

$$\|A_n - H^{-1}\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\gamma}}\right) p.s. \quad et \quad \|A_{n,\tau} - H^{-1}\|^2 = o\left(\frac{\ln n^{1+\delta}}{n}\right) p.s.$$

*De plus, les estimateurs* $\theta_{n,\tau'}$ *définies par* (5) *vérifient*

$$\sqrt{n}(\theta_{n,\tau'} - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, H^{-1}\Sigma H^{-1}),$$

*où* $\Sigma = \mathbb{E}\left[\nabla_h g(X,\theta)\nabla_h g(X,\theta)^T\right]$.

Les estimateurs de Newton Stochastique Universel Moyenné Pondéré atteignent ainsi l'efficacité asymptotique sous des hypothèses très faibles.

# Bibliographie


Bercu, B., Godichon, A., & Portier, B. (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. SIAM Journal on Control and Optimization, 58(1), 348-367.

Bercu, B., Bigot, J., Gadat, S., & Siviero, E. (2023). A stochastic Gauss–Newton algorithm for regularized semi-discrete optimal transport. Information and Inference: A Journal of the IMA, 12(1), 390-447.

Boyer, C., & Godichon-Baggioni, A. (2023). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. Computational Optimization and Applications, 84(3), 921-972.

Duflo, M. (1996). Algorithmes stochastiques (Vol. 23, pp. xiv+-319). Berlin: Springer.

Gadat, S., & Panloup, F. (2017). Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. arXiv preprint arXiv:1709.03342.

Godichon-Baggioni, A. (2019). Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. Journal of Statistical Planning and Inference, 203, 1-19.

Godichon-Baggioni, A., Lu, W., & Portier, B. (2024). Recursive ridge regression using second-order stochastic algorithms. Computational Statistics & Data Analysis, 190, 107854.

Godichon-Baggioni, A., & Lu, W. (2023). Online stochastic Newton methods for estimating the geometric median and applications. arXiv preprint arXiv:2304.00770.

Leluc, R., & Portier, F. (2023). Asymptotic Analysis of Conditioned Stochastic Gradient Descent. Transactions on Machine Learning Research.

Mokkadem, A., & Pelletier, M. (2011). A generalization of the averaging procedure: The use of two-time-scale algorithms. SIAM Journal on Control and Optimization, 49(4), 1523-1543.

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. SIAM Journal on Control and Optimization, 39(1), 49-72.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, 400-407.

# Boosting in Online Non-parametric Regression

Paul Liautaud[1], Pierre Gaillard[2] & Olivier Wintenberger[1,3]

[1] *Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation (LPSM), F-75005 Paris, France ;* {paul.liautaud,olivier.wintenberger}@sorbonne-universite.fr

[2] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France ;* pierre.gaillard@inria.fr

[3] *Wolfgang Pauli Institut, c/o Fakultät für Mathematik, Universität Wien, 1090 Vienna, Austria*

**Résumé.** Dans de nombreuses applications, les données ne sont pas disponibles dès le départ pour apprendre un modèle, mais elles sont observées séquentiellement sous forme de flux de données. De plus, l'environnement est parfois si complexe qu'il est difficile sinon impossible de déterminer un modèle convenable et d'utiliser les techniques d'apprentissage statistique classiques. En particulier les hypothèses d'indépendance ou i.i.d. sur nos données peuvent ne plus être pertinentes. Il est ainsi nécessaire d'adopter une approche robuste en utilisant une méthode qui apprend au fur et à mesure, en tirant des enseignements des données au cours du temps. Tels sont les objectifs de la théorie de l'apprentissage en ligne.

D'autre part, le boosting est une puissante technique d'apprentissage par ensemble introduite par Freund et al. [4] ayant gagné en popularité dans les environnements d'apprentissage batch automatique. Son adaptation au paradigme d'apprentissage séquentiel présente alors des défis et des opportunités uniques. Des efforts récents ont notamment cherché à étendre l'efficacité des algorithmes de boosting à de tels contextes par exemple dans Beygelzimer et al. [1] ou dans Brukhim and Hazan [2]. En entraînant et en optimisant dynamiquement des weak learners (par exemple, de simples arbres de prédiction) sur des données collectées de manière séquentielle, le boosting en ligne permet d'améliorer les performances prédictives et l'adaptabilité aux distributions de données évolutives.

Nous considérerons ici le cadre de la régression non paramétrique séquentielle avec des paires de données $(x_t, y_t)$ arbitraires (comme dans Rakhlin and Sridharan [7] ou Cesa-Bianchi and Lugosi [3]). En particulier, nous analyserons un algorithme de Boosting, en discutant de ses garanties de convergence et en les comparant aux taux optimaux (minimax) déjà établis sous certaines hypothèses par exemple dans Rakhlin and Sridharan [7].

**Abstract.** The rapid proliferation of data streams, coupled with the escalating complexity of data, has precipitated a shift towards sequential methods capable of processing information in real-time. As a consequence, traditional statistical assumptions like stationarity or independently and identically distributed data are no longer tenable. In this context, designing algorithms that can adapt to evolving data streams with minimal assumptions is imperative.

On the other hand, Boosting, a powerful ensemble learning technique presented in Freund et al. [4], has gained significant traction in offline machine learning settings. However, its adaptation to the sequential learning paradigm presents unique challenges and opportunities. Recent efforts have sought to extend the efficacy of boosting algorithms to online learning paradigms for instance in Beygelzimer et al. [1] or Brukhim and Hazan [2]. By dynamically training and optimizing weak learners (e.g. simple decision trees) based on sequentially collected data, online boosting holds promise for enhancing predictive performance and adaptability to evolving data distributions.

Here, we will consider the framework of sequential non-parametric regression with arbitrary data pairs $(x_t, y_t)$ (as in Rakhlin and Sridharan [7] or Cesa-Bianchi and Lugosi [3]). In particular, we will analyze a Boosting algorithm, discussing its convergence guarantees compared to the optimal rates (minimax) already established, for example, in Rakhlin and Sridharan [7].

**Key words.** Statistical Learning, Online Learning, Boosting, Ensemble Learning, Nonparametric Regression.

# 1 Online Nonparametric Regression

## 1.1 Setting

Online nonparametric regression refers to a learning framework where a model is trained sequentially on streaming data to predict an output variable without assuming any specific functional form for the underlying relationship between inputs and outputs. Unlike traditional regression methods, which estimate parameters of a predefined model using all available data at once, online nonparametric regression updates the model continuously as new data points arrive. This adaptive learning process allows the model to capture complex and evolving patterns in the data without requiring strong assumptions about its structure. Online nonparametric regression is particularly well-suited for applications where data is generated sequentially or where the underlying relationship between variables may change over time.

We consider that pairs of data $(x_1, y_1), \ldots, (x_t, y_t), \cdots \in \mathcal{X} \times \mathcal{Y}$ arrive in a stream and we are tasked with sequentially predicting each next response $y_t$ given the current $x_t$ and the past data $\{(x_s, y_s)\}_{s=1}^{t-1}$. Let $\hat{y}_t$ be our prediction and let the quality of this forecast be evaluated via $\ell_t$ (e.g. the square loss $\ell_t(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$). More formally, the scenario can be formalized as follows:

**Learning Scheme**

For each round $t = 1, \ldots, T$, the learner or algorithm

- observes an input $x_t \in \mathcal{X}$
- makes a prediction $\hat{y}_t \in \mathcal{Y}$
- observes the true output $y_t \in \mathcal{Y}$
- measures loss by $\ell_t(\hat{y}_t, y_t)$
- updates his prediction rule

## 1.2 How to make predictions with weak learners?

For each $t = 1, \ldots, T$, we build our predictions by combining $K \geq 1$ *sequential* predictors $(f_{k,t})_t, k \in \{1, \ldots, K\}$ from a class of *weak learners*

$$\mathcal{W} = \{x \mapsto f(x; \theta, I) : \theta \text{ parameter of } f \text{ with support } I\} \subset \mathcal{Y}^{\mathcal{X}},$$

e.g. the set of regression trees with (low) depth 1, $\mathcal{W}_1 = \{f(\cdot; \theta, I) : \theta \in \mathbb{R}^2 \text{ and } I = (I^{(1)}, I^{(2)}), I^{(1)} \sqcup I^{(2)} = \mathcal{X}\}$.

Let $K \geq 1$ and $(f_1, \ldots, f_K) \in \mathcal{W}^K$. A prediction of order $K$ (i.e. using $f_1, \ldots, f_K$) will be, at any time $t \geq 1$,

$$\hat{y}_t = F_{K,t}(x_t) = \sum_{k=1}^{K} f_{k,t}(x_t),$$

using the *strong estimator* $F_K = \sum_{k=1}^{K} f_k$ belonging to the class

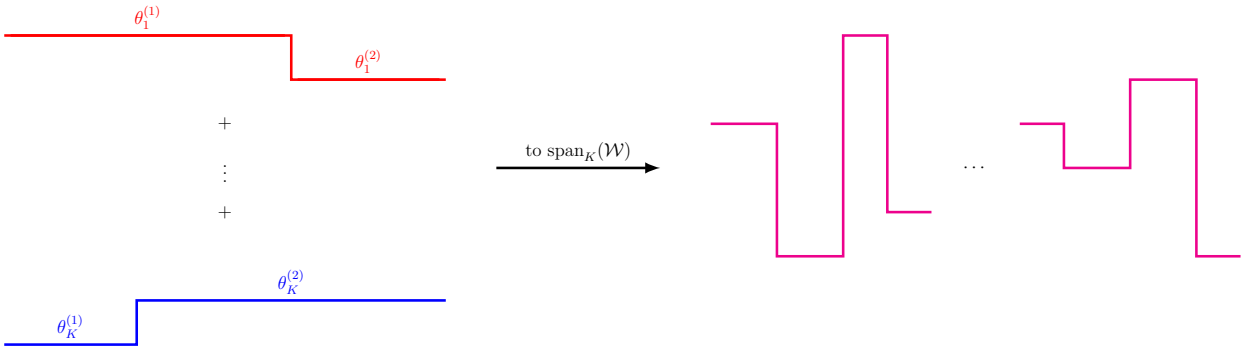
$$\text{span}_K(\mathcal{W}) := \left\{ F_K = \sum_{k=1}^{K} f_k : f_k \in \mathcal{W} \right\}.$$

Why several $f_k$? What is their goal? A weak learner, as the name suggests, is too weak to make accurate predictions on its own. Therefore, it needs to rely on its peers. Considering a mixture of $K$ weak learners, each estimator $f_k$ favors learning from the errors (alias residuals) of $K-1$ others and ensures to correct what it can. The collective learning of these weak learners leads to the formation of a *strong learner* capable of making high-quality predictions.

Example: Combining $K$ predictors in $\mathcal{W}_1 = \{\text{uniform regression trees of depth 1}\}$ which is the set of 2-piecewise constant functions with random cut will lead to a strong predictor in

$$\text{span}_K(\mathcal{W}) \subset \left\{ F_{K'} : F_{K'}(x) = \sum_{k=1}^{K'} \theta_k \mathbf{1}_{x \in J_k}, \theta_k \in \mathbb{R}, \bigsqcup_{k=1}^{K'} J_k = \mathcal{X} \right\} =: \mathcal{F}_{K'},$$

where $\mathcal{F}_{K'}$ is the space of functions that are constant on $K' = 2K + 1$ intervals. Here are some illustrations:



## 1.3 How to measure the performance?

In traditional statistical (and batch) learning theory, the goal is to build a predictor or estimator $\hat{f}_{1:T}$ using a full batch of $T$ data and which minimizes the empirical risk

$$\frac{1}{T}\sum_{t=1}^{T}\ell_t(\hat{f}_{1:T}(x_t), y_t).$$

First problem is that we do not have access to the whole dataset to build such a $\hat{f}_{1:T}$. Second, if the environment chooses large losses $\ell_t$ for all decisions and time $t$, it is then impossible for the learner to ensure a small cumulative loss. Therefore, one needs a relative criterion: the **regret** of the learner defined as the difference between the cumulative loss he incurred and that of the best fixed decision in hindsight.
For a given time horizon $T \geq 1$, and a sequence of losses $(\ell_t)$ the problem of regression is then formulated as that of minimizing the regret

$$\mathrm{Reg}_T(\mathcal{F}) := \sum_{t=1}^{T}\ell_t(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell_t(f(x_t), y_t) \tag{1}$$

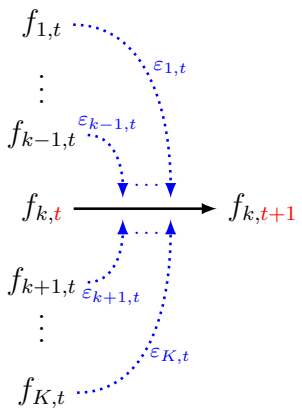
with respect to some benchmark class of non-parametric functions $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ (e.g. the space of Lipschitz functions).
We assume that the losses $(\ell_t)$ are differentiable and convex in their first argument (in $\hat{y}_t$). The problem is now to optimize in each predictor $f_k$, so we can rewrite $\ell_t : \mathrm{span}_K(\mathcal{W}) \times \mathbb{R} \to \mathbb{R}$ as a function $\ell_t : \mathcal{W}^K \to \mathbb{R}$ such that

$$\ell_t(f_{1,t}, \ldots, f_{K,t}) = \ell_t\left(\sum_{k=1}^{K} f_k(x_t), y_t\right).$$

# 2 Analysis of an Online Boosting Algorithm

## 2.1 Architecture of the Algorithm

Let's begin by introducing the algorithm designed to address this problem:




**Algorithm 1: Online Gradient Boosting**

**for** $t = 1$ **to** $T$ **do**
  Receive data $x_t$;
  Predict $\hat{y}_t = \hat{F}_{K,t}(x_t) = \sum_{k=1}^{K} \hat{f}_{k,t}(x_t)$;
  Reveal residuals $\varepsilon_{k,t}$, gradients $g_{k,t} = \nabla_{\hat{f}_{k,t}} \ell_t(\hat{f}_{1,t}, \ldots, \hat{f}_{K,t})$ for all $k = 1, \ldots, K$ and incur $\ell_t(\hat{y}_t, y_t)$;
  **for** $k = 1$ **to** $K$ **do**

$$\hat{f}_{k,t+1} \leftarrow \texttt{gradient\_step}(\varepsilon_{k,t}, g_{k,t}) \tag{2}$$

**Return:** $\hat{F}_{K,T} = \sum_{k=1}^{K} f_{K,t}$

$\varepsilon_{k,t}$ are called the *residuals* or *errors* of each weak learner $f_{k,t}$ at time $t$. At any time $t = 1, \ldots, T$, each weak learner $f_{k,t}$ aims to best correct the *residuals* $\varepsilon_{1,t}, \ldots, \varepsilon_{K,t}$ of $\{f_{1,t}, \ldots, f_{K,t}\} \backslash \{f_{k,t}\}$.

## 2.2 How to bound the regret (1)?

We can decompose the regret (1) as a sum of the following 2 stage regrets:

$$R_T^{(1)} = \sum_{t=1}^{T} \ell_t(\hat{f}_{1,t}, \ldots, \hat{f}_{K,t}) - \inf_{f_1, \ldots, f_K \in \mathcal{W}} \sum_{t=1}^{T} \ell_t(f_1, \ldots, f_K) \tag{3}$$

$$R_T^{(2)} = \inf_{f_1, \ldots, f_K \in \mathcal{W}} \sum_{t=1}^{T} \ell_t(f_1, \ldots, f_K) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_t(f(x_t), y_t) \tag{4}$$

Assume we consider an online version of gradient descent, *online gradient descent (OGD)* (introduced in [8], and which can be applied to any convex and differentiable loss function), to perform (2) sequentially for any $k = 1, \ldots, K$ and any $t \geq 1$ as

$$\hat{f}_{k,t+1} \leftarrow f_{k,t} - \eta_{k,t} g_{k,t} \tag{5}$$

Algorithm (1) verifies the following result for bounding $R_T^{(1)}$:

**Theorem 2.1.** *Assume losses $\ell_t : \mathbb{R}^K \to \mathbb{R}$ are differentiable and $\sigma_k$-strongly convex in each coordinate $k = 1, \dots, K$. Also assume that for any $k = 1, \dots, K, \sup_t |\nabla_{\hat{\theta}_{k,t}} \ell_t| \leq G_k$. Then, Algorithm 1 with (5) and $\eta_{k,t} = \frac{1}{\sigma_k t}$ has the following regret:*

$$\sum_{t=1}^{T} \ell_t(\hat{f}_{1,t}, \dots, \hat{f}_{K,t}) - \inf_{f_1,\dots,f_K \in \mathcal{W}} \sum_{t=1}^{T} \ell_t(f_1, \dots, f_K) \leq \sum_{k=1}^{K} \frac{G_k^2}{2\sigma_k} \ln(T_k + 1) \leq \frac{G^2}{2\sigma} K \ln(T + 1) \quad (6)$$

*with $T_k = |\{t : x_t \in I_k\}|, G = \sup_{1 \leq k \leq K} G_k$ and $\sigma = \inf_{1 \leq k \leq K} \sigma_k$.*

Additionally, we will explore other minimization procedure step (2) within Algorithm 1.

Remember that our goal is to establish a bound for regret (1), necessitating an analysis of the second regret outlined in (4). Specifically, we introduce a *boosting/learning condition* for weak learners in $\mathcal{W}$ with respect to an appropriate class of functions $\mathcal{G}$ (e.g. the set of piecewise constant functions for weak learners in $\mathcal{W}_1$).

**Assumption 2.1.** *For any $k \geq 1, \exists \rho_k = \rho_k(\mathcal{W}, \mathcal{G}) \in [0, 1]$ such that*

$$\inf_{f_1,\dots,f_{k+1} \in \mathcal{W}} \sum_{t=1}^{T} \ell_t(f_1, \dots, f_k, f_{k+1}) - \inf_{g \in \mathcal{G}} \sum_{t=1}^{T} \ell_t(g(x_t), y_t) \leq \rho_k \left( \inf_{f_1,\dots,f_k \in \mathcal{W}} \sum_{t=1}^{T} \ell_t(f_1, \dots, f_k) - \inf_{g \in \mathcal{G}} \sum_{t=1}^{T} \ell_t(g(x_t), y_t) \right) \quad (7)$$

The latter stems from a common assumption prevalent in batch scenarios, notably introduced and scrutinized by Jiang [6] within both regression and classification contexts.

This hypothesis will serve us in managing the second term of regret $R_T^{(2)}$ (4) recursively as we advance through $k$. Following this, it becomes necessary to break down $R_T^{(2)}$ into a regret against a suitable class of functions $\mathcal{G}$ (e.g. $\mathcal{G} = \mathcal{F}_{K'}$), ensuring the existence of a learning coefficient $\rho(\mathcal{W}, \mathcal{G})$.

# 3 Experiments

Consider the standard regression problem,

$$\forall 1 \leq t \leq T, \quad y_t = g(x_t) + W_t$$

where $W_t \sim \mathcal{N}(0, \sigma^2), \sigma > 0, g : \mathbb{R} \to \mathbb{R}$ and $\mathcal{X} = [0, 1]$.

In experiments below, we aim to reconstruct $g : x \mapsto \cos(3\pi x) - \sin(3x)$ using $K = 10$ weak learners with random supports (of type given in the example) and we define $\ell_t$ as the square loss function.
Recall that $\mathcal{F}_{K'}$ represents the set of functions that remain constant across $K' = 2K+1$ intervals. Below, we depict the progression of $\mathrm{Reg}_T(\mathcal{F}_{K'})$ over time, indicating the regret of our algorithm employing regression trees compared to the optimal piecewise constant functions of similar order. Additionally, we present in a separate graph the final prediction generated by our algorithm, juxtaposed with that of its competitor.

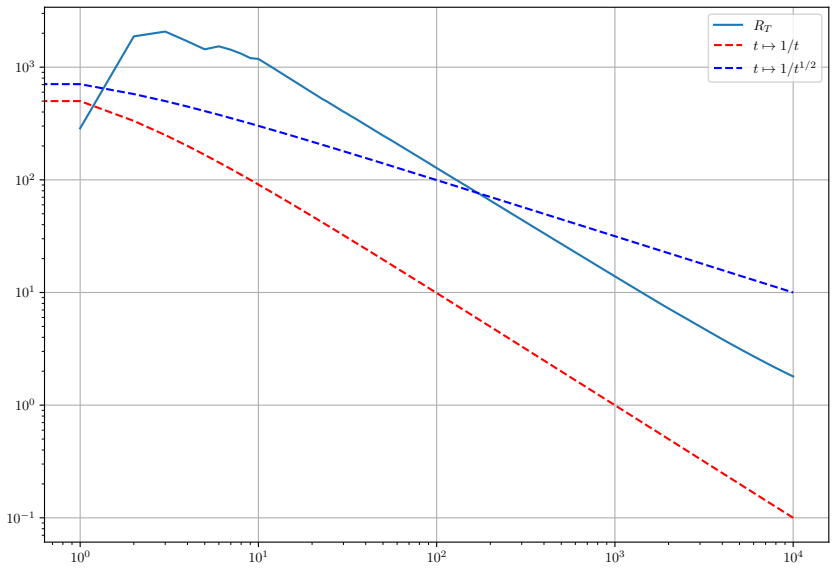



Figure 1: $\mathrm{Reg}_T(\mathcal{F}_{K'})/T$ as a function of $T$.

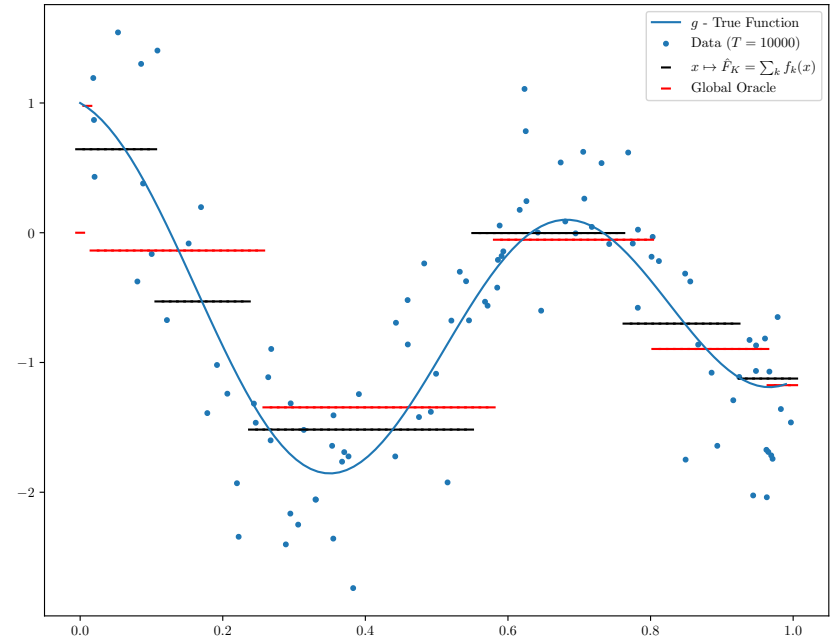



Figure 2: Final prediction

# References


[1] A. Beygelzimer, E. Hazan, S. Kale, and H. Luo. Online gradient boosting. *Advances in neural information processing systems*, 28, 2015.

[2] N. Brukhim and E. Hazan. Online boosting with bandit feedback. In *Algorithmic Learning Theory*, pages 397–420. PMLR, 2021.

[3] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[4] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[5] P. Gaillard and S. Gerchinovitz. A chaining algorithm for online nonparametric regression. In *Conference on Learning Theory*, pages 764–796. PMLR, 2015.

[6] W. Jiang. On weak base hypotheses and their implications for boosting regression and classification. *The Annals of Statistics*, 30(1):51–73, 2002.

[7] A. Rakhlin and K. Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

[8] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

# Statistique appliquée à la Gestion

# A semiparametric location-scale model with application to credit risk

Guillaume Flament[1] & Valentin Patilea[2]

[1] *CREST, Ensai, Univ. Rennes, and Square Research Center ; France, guillaume.flament@square-management.com*
[2] *CREST, Ensai, Univ. Rennes ; France, valentin.patilea@ensai.fr*

**Résumé.** Dans le contexte de la gestion des risques, les institutions financières ont tendance à utiliser des modèles dits "réglementaires", comme par exemple le modèle Merton-Vašíček pour estimer la perte inattendue d'un portefeuille de crédits dans le cas des stress-tests. Dans ce modèle, un facteur commun Gaussien représente l'état de l'économie. Malheureusement, le modèle "réglementaire" ne permet pas de prendre en compte de manière explicite des données macroéconomiques qui permettraient de rafiner les prédictions. Pour ce faire, nous proposons de modéliser le facteur commun à l'aide d'une régression de type "location-scale" et d'estimer ses quantiles conditionnels. Les fonctions "location" et "scale" sont considérées de type semi-paramétriques à direction révélatrice unique, et la loi du terme d'erreur est générale. Plusieurs estimateurs non paramétriques de la fonction de répartition de l'erreur sont proposés. La performance du modèle est illustrée par des simulations. Enfin, une application sur des données réelles issues d'un exercice de stress-test climatique est présentée.

**Mots clefs.** Modèles semi-paramétrique, Lissage à noyau, Économétrie, Finance.

**Abstract.** In the context of risks management, financial institutions tend to use "regulatory models", such for example the Merton-Vašíček model which estimates the unexpected loss of a credit portfolio in the case of credit stress-testing. In this model, a Gaussian common factor represents the state of the economy. This model does not allow to explicitly account for the information from macroeconomic data. We therefore propose to model this common factor with a semiparametric single-index location-scale model for estimating conditional quantiles of the common factor. Several nonparametric estimators of the error distribution function are proposed. The finite sample performance is investigated by a simulation study. Finally, an application on real data from a climate stress-testing exercise is presented.

**Keywords.** Semiparametric models, Kernel smoothing, Econometrics, Finance.

## 1 Introduction

Financial institutions are required to manage their unexpected losses, which refer to financial losses that cannot be accurately forecast using conventional risk management models. Unexpected losses can result from a large range of factors, such as incidents on financial markets (e.g., the subprime crisis), natural disasters (e.g., the tidal waves following earthquakes and the Fukushima nuclear power plant failure), economic crises (e.g., the Asian crisis in the 90's),

counterparty failures, sudden regulatory changes, etc. These losses are typically challenging to predict due to their complex and often unique nature. To prevent financial institutions defaults as consequence of such unexpected losses, for each borrower[1] all banks are required to have a minimum amount of capital (see also, Roncalli, 2020).

This capital requirement depends on the $EAD$ which is *Exposure at Default* and represents the amount of money a lender is exposed to, or stands to lose, if a borrower fails to fulfill their financial obligations. It also depends on the maturity of the loan, the *Losses Given Defaults* (LGD) which is the amount that could not be retrieved by the bank after all assets are liquidated. It finally depends on the probability of default over a given period (e.g., one year). Let $p$ be the unconditional probability of default of a lender.

In the Merton, 1974 model, a company defaults when the normalized value of its assets $A$ falls below a given threshold $B$. Let $D$ be a default indicator such that :

$$\{D = 1\} \iff \{A \leq B\}. \tag{1}$$

Therefore, assuming that $A$ is standard Gaussian random variable, the unconditional probability of default is $p = \mathbb{P}(D = 1) = \Phi(B)$ where $\Phi(\cdot)$ is a standard Gaussian distribution function. Furthermore one can assume that $A$ depends on an idiosyncratic factor $\epsilon$ and a random common factor $Y$ both supposed to be standard Gaussian variables. The common factor is then interpreted as the state of the global economy, the lower its value the worst the state of the economy, and higher the probability of default. This leads to the so-called Merton-Vašíček model, see Vašíček, 2002. In that model, assuming further that the correlation between $A$ and $Y$ is equal to $\sqrt{\xi}$, the random variable $A$ is decomposed as

$$A = \sqrt{\xi}Y + \sqrt{1-\xi}\,\epsilon. \tag{2}$$

Given that $B = \Phi^{-1}(p)$, the conditional probability of default given $Y = y$ can be obtained by plugin (2) in (1) which yields :

$$\pi(y) = \mathbb{P}(\sqrt{\xi}y + \sqrt{1-\xi}\epsilon \leq \Phi^{-1}(p)) = \Phi\left(\frac{\Phi^{-1}(p) - \sqrt{\xi}\,y}{\sqrt{1-\xi}}\right). \tag{3}$$

By construction, we have $p = \mathbb{E}(\pi(Y))$. Moreover, $\pi(y)$ is a decreasing function of $y$. Using the relationship $\Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$, $\alpha \in (0,1)$, and replacing $y$ by minus the $\alpha-$th quantile of the standard Gaussian distribution, we can rewrite (3) and redefine under the form

$$\pi(\alpha) = \Phi\left(\frac{\Phi^{-1}(p) + \sqrt{\xi}\,\Phi^{-1}(\alpha)}{\sqrt{1-\xi}}\right). \tag{4}$$

To prevent banks' failure due to extreme event, the regulator is interested by the evaluation the conditional probabilities corresponding to extreme left quantiles $\Phi^{-1}(\alpha)$. In the case of the Basel II regulation, $\alpha = 0.001$. The required capital due to credit risk for a banking

[1]The regulation is available here : https://www.bis.org/publ/bcbs107.pdf. Formulae for the Internal Rating based, or Foundation Internal Rating Based Ratings (FIRB), method are stated in part 2 section 3, paragraph 272.

institution is then depending on $\pi(\alpha)$. The higher the probability of default, the higher the required capital.

The classical Merton-Vašíček model presents several pitfalls. The most important one is that it assumes that the required capital does not directly depends on any covariates. Indeed, the regulatory formula (3) simply evaluates the conditional probability of default given $Y = y$, without any reference to observed covariates. In the short term, this is perhaps less problematic for the calculation of the required capital. Indeed, fixing this required capital under the scenario that a financial crisis will occur in the next year seems reasonable, since stress-testing exercises aim at protecting the bank from a catastrophic event. However, in the long run, for example for climate stress-testing exercises, the practitioners would prefer to take into account the dynamics of macroeconomic, financial and other types of predictors.

Therefore we construct a statistical model that links the common factor $Y$ to a vector of covariates $W$ (e.g., macroeconomic variables, financial variables). More precisely, our goal will be to estimate the conditional distribution function of $Y$ given the covariates $W$. We will then be able to estimate the conditional quantiles of $Y$ given $W$. In particular, this will allow practitioners to include covariates in the regulatory stress-testing exercise by simply replacing the unconditional (marginal) quantiles $\Phi^{-1}(\alpha)$ by their conditional versions in (4).

## 1.1 A brief look at the statistical model

Let $\tau \in (0, 1)$ and define the conditional quantile function of order $\tau$

$$q_\tau(w) = \inf\{y \ : \Psi_{Y|W}(y \mid w) \geq \tau\} \qquad \text{where} \qquad \Psi_{Y|W}(y \mid w) = \mathbb{P}(Y \leq y \mid W = w).$$

Here, $\Psi_{Y|W}$ is the conditional cumulative distribution function of $Y$ given the covariate vector value, and we have

$$\Phi_Y(\cdot) = \mathbb{E}[\Psi_{Y|W}(\cdot \mid W)],$$

where $\Phi_Y$ is the unconditional (marginal) distribution of $Y$. Under the Gaussianity assumptions like in Merton-Vašíček's model, $\Phi_Y = \Phi$ and the conditional version of (3) given $W = w$ is

$$\pi(\tau \mid w) = \Phi\left(\frac{\Phi^{-1}(p) + \sqrt{\xi}\, q_\tau(w)}{\sqrt{1-\xi}}\right), \qquad \tau \in (0, 1). \tag{5}$$

In our statistical analysis, $\tau$ can be any fixed value between 0 and 1. For the bank regulation rules, for example $\tau = \alpha = 0.001$.

We here consider that the value $p$ is given. In the credit risk analysis, the probability of default $p$ can be accurately estimated from external data, such as the so-called transition matrices obtained from the reports of the rating agencies.

A flexible model for the conditional distribution of $Y$ given $W$, and thus for the conditional quantile function, can be constructed as a location-scale regression model

$$Y = m(Z) + \sigma(X)\varepsilon, \tag{6}$$

where $Z \in \mathbb{R}^{d_Z}$ and $X \in \mathbb{R}^{d_X}$ are subvectors of $W \in \mathbb{R}^{d_W}$, with possibly common components in which case $d_Z + d_X < d_W$. The error term $\varepsilon$ is independent of $W$, and the functions $m$

and $\sigma$ are unknown. For identification purposes, the variance $\mathrm{Var}(\varepsilon)$ has to be set to some value. The location-scale model was extensively studied in the statistically literature. See, for example, Akritas and Van Keilegom, 2001, Neumeyer and Van Keilegom, 2010, Racine and Van Keilegom, 2020. In the location-scale model, the conditional quantile function is

$$q_\tau(w) = \inf_y \left\{ y \ : F_\varepsilon \left( \frac{y - m(z)}{\sigma(x)} \right) \geq \tau \right\}, \tag{7}$$

where $F_\varepsilon(\cdot)$ denotes the distribution function of $\varepsilon$. The vectors $w$, $z$ and $x$ belong to the support of $W$, $Z$ and $X$, respectively.

In order to avoid the curse of dimensionality due to the nonparametric estimation of the multivariate location and scale functions $m$ and $\sigma$, we here consider a single-index modeling of these functions. Under the single-index assumptions, the model (6) becomes

$$Y = m(Z^\top \gamma) + \sigma(X^\top \beta)\varepsilon, \tag{8}$$

where $\gamma$ and $\beta$ are unknown vectors and $m(\cdot)$ and $\sigma(\cdot)$ are now univariate function to be estimated. See also Neumeyer and Van Keilegom, 2010.

In this paper, we study the semiparametric single-index location-scale model (8). Moreover, we consider a smooth estimator of $F_\varepsilon(\cdot)$ using the residuals of the single-index location scale model, following the lines of Azzalini, 1981, see also Neumeyer and Van Keilegom, 2010 and Racine and Van Keilegom, 2020. Keeping in mind the application to credit risk and stress-testing, we consider that the marginal distribution of the response $Y$ is known (for example is the standard normal). This introduces some specific identification constraints in the treatment of the model (8). Knowing the marginal distribution of the response also allows us to propose a data-driven rule for the smooth estimates of the error distribution function. The paper is organized as follows. In section 2 we introduce the single-index estimators of the location and scale functions. Moreover, we consider the estimator of the error distribution function $F_\varepsilon$ obtained as the empirical distribution function of the residuals of the semiparametric location-scale model (8). For the purpose of estimating extreme conditional quantiles, we also consider smoothed versions of the empirical distribution function of the residuals. The finite sample performance of our estimators is investigated by simulations, the results are presented in section 3.1. A real data application is presented in section 3.2.

# 2 Model estimation

The goal is to estimate the finite-dimensional parameters $\gamma$ and $\beta$, the univariate functions $m(\cdot)$ and $\sigma(\cdot)$ and the distribution $F_\varepsilon$ of the error term $\varepsilon$. Plugging into (7) these estimates yields the estimate of $q_\tau(w)$. Let $(Y_i, W_i)$, $1 \leq i \leq n$, be an independent sample of $(Y, W) \in \mathbb{R} \times \mathbb{R}^{d_W}$. Thus $Z_i$ and $X_i$ are independent copies of $Z$ and $X$, the subvectors of $W$ appearing in (8).

## 2.1 Single-index estimators

In the first step, we estimate $\gamma$ and $m$ by semiparametric least squares. By construction, we have

$$\mathbb{E}[Y \mid W] = \mathbb{E}[Y \mid Z^\top \gamma], \qquad \text{for some } \gamma \in \mathbb{R}^{d_Z}.$$

For identification purposes, we set a component $\gamma$ (e.g., the first one) equal to 1. Next, given a value $\gamma$, we consider the regression function

$$m(t;\gamma) = \mathbb{E}[Y \mid Z^\top \gamma = t], \qquad t \in \mathbb{R}.$$

This function can be estimated by local linear smoothing (Fan and Gijbels, 1996), that is

$$\widehat{m}(t;\gamma) = \arg\min_{a} \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} \left\{Y_i - a - b(Z_i^\top \gamma - t)\right\}^2 k\left(\frac{Z_i^\top \gamma - t}{h}\right), \tag{9}$$

where $k$ is a second order symmetric kernel and $h$ is the bandwidth. The index $\gamma$ is then defined as the least squares estimator

$$\widehat{\gamma} = \arg\min_{\gamma} \sum_{i=1}^{n} \left\{Y_i - \widehat{m}(Z_i\gamma^\top;\gamma)\right\}^2, \tag{10}$$

where the optimization is considered under the constraint that the first component of $\gamma$ is equal to 1. See, for example, Ichimura, 1993, Hardle et al., 1993, Carroll et al., 1997, Delecroix et al., 2006 for some references on semiparametric single index models.

To estimate the scale factor $\sigma(\cdot)$ under the single-index assumption, we set the first component of $\beta$ equal to 1, and define

$$\sigma^2(t;\beta) = \frac{1}{\mathrm{Var}(\varepsilon)} \mathbb{E}\left[\left\{Y - m(Z^\top\gamma;\gamma)\right\}^2 \mid X^\top\beta = t\right], \qquad t \in \mathbb{R}.$$

In usual location scale models, the variance of $\varepsilon$ is set equal to 1, and hence no scaling by the inverse of $\mathrm{Var}(\varepsilon)$ is required. However, for the application we have in mind for stress-testing, where there is an additional information on the unconditional distribution of $Y$ (typically, $Y$ is standard Gaussian), the usual identification choice $\mathrm{Var}(\varepsilon) = 1$ is no longer admissible. We therefore have to set a specific identification constraint for the scale function, such for instance

$$\mathrm{Var}(\varepsilon) = \sigma_Y^2/2 := \mathrm{Var}(Y)/2.$$

To estimate $\sigma^2(t;\beta)$ we consider the local linear smoother

$$\widehat{\sigma}^2(t;\beta) = \arg\min_{a} \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} \left\{(2/\sigma_Y^2)\left[Y_i - \widehat{m}(Z_i^\top\widehat{\gamma};\widehat{\gamma})\right]^2 - a - b(X_i^\top\beta - t)\right\}^2 k\left(\frac{X_i^\top\beta - t}{h}\right). \tag{11}$$

For simplicity we use the same bandwidth rate for the local linear estimators of $m(\cdot;\gamma)$ and $\sigma^2(\cdot;\beta)$. The semiparametric least squares estimator of $\beta$ is then

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left[(2/\sigma_Y^2)\left\{Y_i - \widehat{m}(Z_i^\top\widehat{\gamma};\widehat{\gamma})\right\}^2 - \widehat{\sigma}^2(X_i^\top\beta;\beta)\right]^2, \tag{12}$$

the minimization being considered under the identification assumption that the first component of $\beta$ is equal to 1. See Zhu et al., 2013; see also Fan and Yao, 1998, Yin et al., 2010. Let us point out that, the same $\widehat{\beta}$ can be obtained without the scaling factor $(2/\sigma_Y^2)$ in (11) and (12). However, the scaling is necessary for the construction of the residuals and the estimation of the distribution function, as explained below.

## 2.2 Error distribution function estimators

Given the estimators $\widehat{m}(Z_i^\top\widehat{\gamma};\widehat{\gamma})$ and $\widehat{\sigma}^2(X_i^\top\widehat{\beta};\widehat{\beta})$ obtained from (9), (10) and (11), (12), respectively, we define the residuals

$$\widehat{\varepsilon}_i = \frac{Y_t - \widehat{m}(Z_i^\top\widehat{\gamma};\widehat{\gamma})}{\widehat{\sigma}(X_i^\top\widehat{\beta};\widehat{\beta})}, \qquad 1 \leq i \leq n.$$

Based on the residuals $\widehat{\varepsilon}_i$, we study four methods to estimate the distribution function $F_\varepsilon$.

The first method is using the empirical distribution function of the residuals :

$$\widehat{F}_{\varepsilon,\text{emp}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\widehat{\varepsilon}_i \leq t\}}, \qquad t \in \mathbb{R}. \tag{13}$$

See Koul et al., 2017 for a similar estimator in the case of a single-index model for the location function, and without scale function. See also Neumeyer and Van Keilegom, 2010 for a general setup of location scale regression.

This empirical distribution may be a simple and quite effective estimator of $F_\varepsilon(\cdot)$ if the interest lies in quantiles $q_\tau(\cdot)$ with $\tau$ away from 0 and 1. It is however expected to behave poorly for extreme quantiles. We therefore consider smoothed versions of the empirical distribution. More precisely, for some kernel function $k(\cdot)$, which may not identical to that in (9), (11) or (12), we consider

$$\widehat{F}_{\epsilon,1}(u) = \frac{1}{n}\sum_{i=1}^{n} K\left((\widehat{\varepsilon}_i - u)/h_\varepsilon\right) \qquad \text{with} \quad K(u) = \int_{-\infty}^{u} k(v)dv, \tag{14}$$

and $h_\varepsilon$ some bandwidth devoted to the estimation of $F_\varepsilon$. In the case of the Epanechnikov kernel, that is with $k(u) = (3/4)(1-u^2)\mathbb{1}_{\{|u|\leq 1\}}$, we get

$$K(u) = (1/4)\left(3u - u^3 + 2\right)\mathbb{1}_{\{|u|\leq 1\}} + \mathbb{1}_{\{u\geq 1\}}.$$

The smooth empirical distribution is the distribution function obtained by integrating the Parzen-Rosenblatt density estimator. See Azzalini, 1981 for the properties of the smooth empirical distribution estimator in the case where the sample of the variable (here $\varepsilon$) is observed. See also Racine and Van Keilegom, 2020 for the case where $\widehat{\varepsilon}_i$ are obtained after fitting a nonparametric location scale model.

Finally, keeping in mind the application to stress-testing where the interest focuses on the accurate estimation of the left tail of the distribution, we also investigate a version of the smoothed empirical distribution with asymmetric kernel. More precisely, we replace $k(u)$

by $k_l(u) = 2k(u)\mathbb{1}_{\{u\leq 0\}}$, and thus $K(u)$ by $K_l(u) = 2\int_{-\infty}^{u} k(v)dv$, and construct a third estimator under the form

$$\widehat{F}_{\epsilon,2}(u) = \frac{1}{n}\sum_{i=1}^{n}\left\{2K\left((\widehat{\varepsilon}_i - u)/h_\varepsilon\right)\mathbb{1}_{\{\widehat{\varepsilon}_i\leq u\}} + \mathbb{1}_{\{\widehat{\varepsilon}_i\geq u\}}\right\}. \tag{15}$$

In our empirical study, we consider this idea with $k(\cdot)$ the Epanechnikov kernel.

Finally, in order to diminish the bias of the smoothed empirical distribution estimator, we consider it with a higher order kernel. For simplicity, we only investigate the case of the asymmetric kernel. More precisely, we define

$$\widehat{F}_{\epsilon,3}(u) = \frac{1}{n}\sum_{i=1}^{n}\left\{2K_6\left((\widehat{\varepsilon}_i - u)/h_\varepsilon\right)\mathbb{1}_{\{\widehat{\varepsilon}_i\leq u\}} + \mathbb{1}_{\{\widehat{\varepsilon}_i\geq u\}}\right\}, \tag{16}$$

with

$$K_6(u) = C_1\left[u - 7u^3/3 + 63u^5/25 - 33u^7/25 - C_2\right]\mathbb{1}_{\{|u|\leq 1\}} + \mathbb{1}_{\{u>1\}},$$

where $C_1 = 525/256$ and $C_2 = 128/525$.

### 2.3 Semiparametric quantile function estimation

Given the semiparametric estimators of the single-index location and scale functions, as well as an estimator $\widehat{F}_\varepsilon(\cdot)$ of the distribution function $F_\varepsilon(\cdot)$ of $\varepsilon$, we can build our semi-parametric estimator of the conditional quantile function. More precisely, we define

$$\widehat{q}_\tau(w) = \widehat{q}_\tau(z,x) = \inf\left\{y \ : \widehat{F}_\varepsilon\left(\frac{y - \widehat{m}(z^\top\widehat{\gamma})}{\widehat{\sigma}(x^\top\widehat{\beta})}\right)\geq\tau\right\}. \tag{17}$$

where $w$, $z$ and $x$ belong to the support of $W$, $Z$ and $X$, respectively. Our estimator requires several tuning parameters. For the estimators $\widehat{m}(Z_i^\top\widehat{\gamma};\widehat{\gamma})$ and $\widehat{\sigma}^2(X_i^\top\widehat{\beta};\widehat{\beta})$ we impose bandwidths in the range required for deriving the $\sqrt{n}-$asymptotic normality of $\widehat{F}_\varepsilon(\cdot)$. When considering the smoothed versions of the empirical distribution of the residuals $\widehat{\varepsilon}_i$, we propose a choice of the bandwidth which use the unconditional distribution of $Y$; see section 3.1 below.

## 3 Empirical study

In the following, we illustrate our semi-parametric approach for the estimation of the conditional quantiles by means of simulations and a real data application.

### 3.1 Simulation study

In this section we study the estimator proposed in (17) using the location scale model

$$Y = \left(Z^\top\gamma\right)^3 + \exp(X^\top\beta)\varepsilon, \tag{18}$$

where $X, Z \in \mathbb{R}^3$ have three components, and they don't have common components. The variable $Y$ has a standard Gaussian distribution. The parameters are fixed as $\gamma = (1, -1.2, 0.4)$ and $\beta = (1, 0.8, -0.2)$. The covariate vectors are decomposed like $X = (X_1, \widetilde{X})$ and $Z = (Z_1, \widetilde{Z})$, and $\widetilde{X}$ and $\widetilde{Z}$ are simulated as standard Gaussian bivariate random vectors. The variables $\varepsilon$, $X_1$ and $Z_1$ are generated such that $Y \sim \mathcal{N}(0, 1)$. This is achieved using the Box-Müller method, see Box and Muller, 1958. We generate iid samples from (18), for sample sizes from $n \in \{50, 100, 150, 250\}$.

The values $\tau \in \{0.01, 0.05, 0.2\}$ are considered. In both local linear problem (9) and (11), we assumed a bandwidth $h = n^{-1/3.5}$ which matches the assumptions needed for the theory. We consider four different methods to estimate the distribution function $F_\varepsilon(\cdot)$. The first one is $\widehat{F}_{\varepsilon,\text{emp}}(\cdot)$ from (13), the second is the smoothed version (14), denoted by $\widehat{F}_{\varepsilon,1}(\cdot)$, and the last ones are asymmetric smoothed estimators as in (15) and (16), denoted by $\widehat{F}_{\varepsilon,2}(\cdot)$ and $\widehat{F}_{\varepsilon,3}(\cdot)$, respectively. For each smooth estimator $\widehat{F}_{\epsilon,j}(\cdot)$ of $F_\varepsilon(\cdot)$, the bandwidth is set as the solution to the following minimization problem :

$$\min_h \sum_{i=1}^{n} \left( \widehat{F}_{\epsilon,j} \left( \frac{q_\tau - \widehat{m}(\widehat{\gamma}^\top Z_i)\}}{\widehat{\sigma}(\widehat{\beta}^\top X_i)} \right) - \tau \right)^2, \qquad j \in \{1, 2, 3\},$$

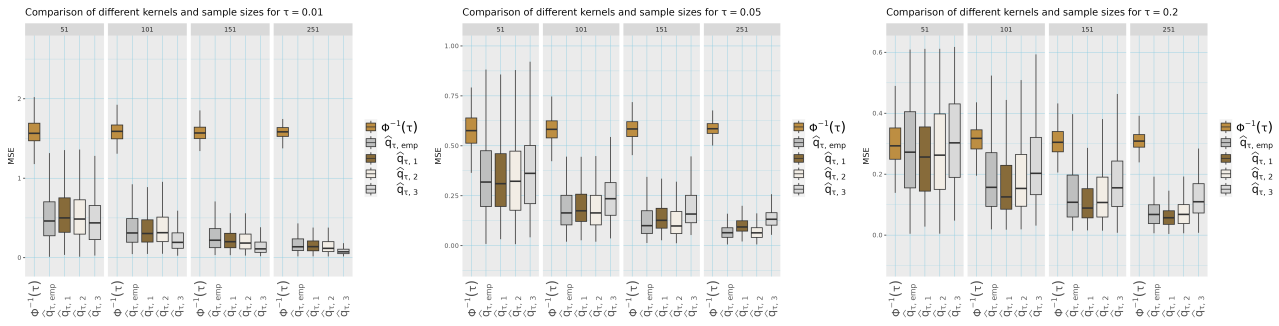
where $q_\tau$ is the marginal quantile of $Y$, here $q_\tau = \Phi^{-1}(\tau)$.

Let us write $\widehat{q}_\tau(Z_i, X_i)$ instead of $\widehat{q}_\tau(W_i)$ when the estimator of the $\tau-$th order conditional quantile function is computed at the observed values of the covariate vector. To measure the accuracy of our estimates of the conditional quantile function, we use the mean square error

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \widehat{q}_\tau(Z_i, X_i) - q_\tau(Z_i, X_i) \right\}^2, \tag{19}$$

with $\widehat{q}_\tau(Z_i, X_i)$ computed according to (17). We also compute the mean squared error with $\Phi^{-1}(\tau)$ replacing $\widehat{q}_\tau(Z_i, X_i)$. For each simulation setup (kernel choices, $\tau$, sample sizes) we report the performance based on $R = 250$ replications.

Figures 1a, 1b, 1c illustrates the mean squared error (19).




(a) (b) (c)

Figure 1: Mean squared error (19) for all proposed kernels, panel (a), (b) and (c) correspond to $\tau = 0.01, 0.05, 0.2$, respectively.

Across all quantiles below a certain threshold , the estimates obtained with $\widehat{F}_{\varepsilon,2}(\cdot)$ and $\widehat{F}_{\varepsilon,3}(\cdot)$ (using asymmetrical kernels) exhibit slightly better performance compared to those obtained from $\widehat{F}_{\varepsilon,\text{emp}}(\cdot)$ and $\widehat{F}_{\varepsilon,1}(\cdot)$ (see Figures 1a and 1b). This is especially true for smaller samples. However, at the highest quantile level considered ($\tau = 0.2$), using the smooth estimator $\widehat{F}_{\varepsilon,1}(\cdot)$ (constructed with a symmetric kernel) for estimating $F_{\varepsilon}(\cdot)$ yields superior performance (see Figure 1c). It is worth noting that all the alternative methods we propose drastically outperforms the regulatory framework.

## 3.2 Use case : Banking climate stress-testing

Let us apply the location-scale model with real data. The goal is to use in (4) the conditional quantiles of the common factor given covariates given by macroeconomic variables. In particular, the goal is to use macroeconomic scenarios that integrates climate risks to obtain forward looking probabilities of default that include this risk. Therefore, we will first estimate $\beta, \gamma$, $m$ and $\sigma$ using past data and following the method proposed in section 2. Then, conditionally given a scenario from the NGFS[2], we compute $\widehat{q}_{\tau}(z, x)$ for covariate values $z$ and $x$ corresponding to each year between 2023 and 2050 available in the macroeconomic scenario published by the NGFS.

Our historic data consists of corporate default rates from the S&P report and economic variables from U.S. Bureau of Economic Analysis, 2023; World Bank, 2023 from 1981 to 2021. Our vector of predictors $W$ include GDP growth and inflation rate, in this case $X = Z = W$ and $x = z = w$ for the NGFS's scenario values. We assume that $\xi$, the correlation between the normalized assets' value and the common factor, is given by the regulator's formula, and $p$ is also given. The Figure 2 represents the estimated conditional distribution of $Y$ given $W = w$, while the Figure 3 represents $\pi(\tau \mid w)$ computed as in (5).

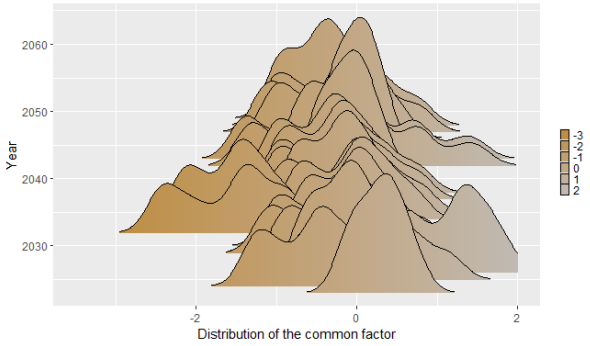



Figure 2: Distribution of the common factor conditionally to the Divergent Net Zero scenarios from the NGFS used during stress-testing exercises.

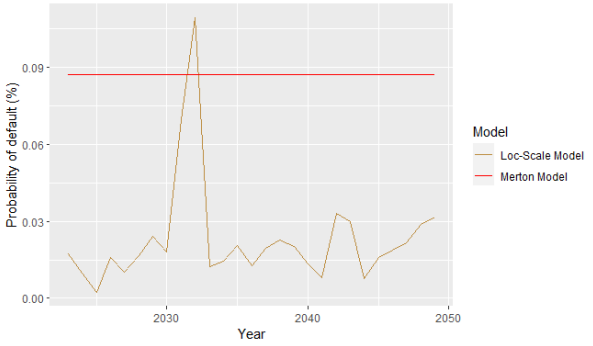



Figure 3: Conditional probability of default (under model (5)) at quantile 1% conditionally to the 'Below 2°C' scenario from the NGFS used during stress-testing exercises.

[2]See https://www.ngfs.net/ngfs-scenarios-portal/. The Divergent Net Zero scenario is used.

# References


Akritas, M. G., & Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, *28*(3), 549–567.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, *68*(1), 326–328.

Box, G. E., & Muller, M. E. (1958). A note on the generation of random normal deviates. *The annals of Mathematical Statistics*, *29*(2), 610–611.

Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.

Delecroix, M., Hristache, M., & Patilea, V. (2006). On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference*, *136*(3), 730–769.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications* (Vol. 66). Chapman & Hall, London.

Fan, J., & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, *85*(3), 645–660.

Hardle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *The annals of Statistics*, *21*(1), 157–178.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, *58*(1-2), 71–120.

Koul, H. L., Müller, U. U., & Schick, A. (2017). Estimating the error distribution in a single-index model. *From Statistics to Mathematical Finance: Festschrift in Honour of Winfried Stute*, 209–233.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, *29*(2), 449–470.

Neumeyer, N., & Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, *101*(5), 1067–1078.

Racine, J. S., & Van Keilegom, I. (2020). A smooth nonparametric, multivariate, mixed-data location-scale test. *Journal of Business & Economic Statistics*, *38*(4), 784–795.

Roncalli, T. (2020). *Handbook of financial risk management*. CRC Press.

U.S. Bureau of Economic Analysis. (2023). Gross domestic product [GDPA] [February 5, 2024.].

Vašíček, O. (2002). The distribution of loan portfolio value. *Risk*, *15*(12), 160–162.

World Bank. (2023). Inflation, consumer prices for the united states [fpcpitotlzgusa] [Accessed February 5, 2024]. https://fred.stlouisfed.org/series/FPCPITOTLZGUSA

Yin, J., Geng, Z., Li, R., & Wang, H. (2010). Nonparametric covariance model. *Statist. Sinica*, *20*(1), 469–479.

Zhu, L., Dong, Y., & Li, R. (2013). Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statist. Sinica*, *23*(3), 1235–1255.

# Modèles probabilistes pour les permutations et dépendances

Arthur Fétiveau [1] & Gilles Durrieu [2] & Emmanuel Frénod [3]

[1] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , arthur.fetiveau@univ-ubs.fr*
[2] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , gilles.durrieu@univ-ubs.fr*
[3] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , emmanuel.frenod@univ-ubs.fr*

**Résumé.** Dans l'objectif d'améliorer la performance financière d'une entreprise, nous devons classer tous les produits qu'elle commercialise selon leurs intérêts en utilisant des critères financiers. Les critères et le classement final pouvant se modéliser sous formes de permutations, nous nous intéressons aux modèles de Mallows qui sont des modèles définis dans l'espace des permutations. Nous nous intéressons notamment à la question de la dépendance dans les modèles de Mallows. Dans cette optique, nous introduisons une approche basée sur une fonction de coût minimisant l'impact de la dépendance entre les critères.

**Mots-clés.** Apprentissage, dépendance, distance de Kendall, modèles de Mallows, statistique computationnelle.

**Abstract.** In order to improve the financial performance of a company, we must classify all their commercialised products according to their interests using financial criteria. The criteria and the final classification can be modeled using permutations. We consider Mallows models defined in the space of permutations. We are particularly interested in the question of dependence in Mallows models. Here, we introduce an approach based on a cost function minimizing the impact of the dependence between criteria.

**Keywords.** Machine learning, dependence, Kendall distance, Mallows models, computational statistics.

## 1 Introduction

Dans une entreprise disposant de nombreux produits, il n'est pas possible d'analyser dans le détail chacun d'entre eux. Il y aurait trop de produits à analyser. Ainsi, il est primordial d'avoir un outil permettant, selon les préférences du décideur, de prioriser les éléments les plus importants pour le décideur. Différentes approches ont été abordées pour faire de l'agrégation multi-critère, telles que la méthode MAUT (Multi Attribute Utility Theory) notamment explorée par Keeney et Raiffa (1976) ou encore les différentes méthodes ELECTRE (ELimination Et Choix Traduisant la REalité), expliquée par Roy (1968), et PROMETHEE (Preference Ranking Organisation METHod for Enrichment Evaluations) de Mareschal, Brans, et Vincke (1984).

Afin de nous aider à décider, nous disposons de critères basiques tels que le chiffre d'affaires ou la prévision de la quantité de produits à vendre. Nous pouvons également créer de nombreux critères dérivés générant des informations différentes comme par exemple l'évolution de l'écart entre la quantité vendue et la quantité prévue de ventes. Pour modéliser notre problème, nous utilisons un modèle de Mallows (Mallows (1957); Fétiveau (2024)) sur l'ensemble des permutations. Cependant dans cette modélisation, les critères sont supposés indépendants entre eux, ce qui n'est pas le cas en pratique. Nous ne pouvons pas non plus décider de n'utiliser que les critères utiles puisque nous n'avons aucune idée de si certains critères dérivés apportent des informations nouvelles. Par conséquent, nous introduisons une fonction de coût à optimiser permettant de trouver les valeurs de $\theta$ dans le cas de variables dépendantes. Nous testons les résultats sur des données simulées suivant le modèle de Mallows.

## 2 Modèle de Mallows et estimation des paramètres

Soit une suite de $n$ éléments à ordonner. On appelle une permutation $\pi$, une bijection de $\{1, ..., n\}$ dans $\{1, ..., n\}$. On note $\pi(i)$ le rang associé à l'élément $i$ et $\pi^{-1}(i)$ l'élément associé au rang $i$ de $\pi$ pour $i \in \{1, ..., n\}$. On note $\mathcal{S}_n$ l'ensemble de toutes les permutations possibles de $n$ éléments. On a $\#\mathcal{S}_n = n!$, où $\#$ est le cardinal.

Le modèle de Mallows (1957) est défini comme une distribution de paramètres $\pi$ et $\theta \in \mathbb{R}$. La probabilité d'une permutation $\sigma$ est donnée par

$$P_\theta(\sigma) = \frac{exp(-\theta\ d(\sigma, \pi))}{Z(\theta)} \tag{1}$$

où $\pi$ est la permutation modale représentant le véritable classement inconnu, $\theta \in \mathbb{R}$ est le paramètre de dispersion autour du classement modal, $d(.,.)$ est une distance invariante à la relabélisation et $Z(\theta)$ le terme de normalisation. Ce dernier terme ne dépend pas de $\pi$ lorsque la distance utilisée possède la propriété d'invariance à la relabélisation. Quand $\theta = 0$, alors chaque permutation $\sigma$ de $\mathcal{S}_n$ est équiprobable. En particulier quand $\theta$ est positif, la probabilité est concentrée autour de la permutation modale et quand $\theta$ est négatif, la probabilité est concentrée autour de la permutation antimodale.

Dans cet article, on utilise la distance de Kendall (1938) donnée par

$$\begin{aligned} d_k(\pi, \sigma) \quad &= \sum_{i=1}^{n-1} \sum_{j>i} \mathbb{1}\left([\pi(i) > \pi(j) \wedge \sigma(i) < \sigma(j)] \vee \right. \\ &\left. [\pi(i) < \pi(j) \wedge \sigma(i) > \sigma(j)]\right), \end{aligned} \tag{2}$$

qui est bien une distance invariante à la relabélisation (Diaconis (1988)). Cette distance bénéficie également d'une propriété importante puisqu'il est possible de la décomposer en une somme finie de termes indépendants. La distance de Kendall détermine le nombre minimal de transpositions adjacentes pour passer d'une permutation à l'autre.

Nous considérons pour nos objectifs que la permutation $\pi$ représente le classement idéal, celui que nous devons trouver par consensus, c'est à dire via un accord le plus satisfaisant pour tous, entre les différents critères. Pour tout $j \in \{1, ..., J\}$, chacun de nos critères est représenté par $\sigma_j$ et est issu du modèle de Mallows de paramètres $\pi$ et $\theta_j$ (1). Pour tout $j \in \{1, ..., J\}$, les estimateurs des paramètres $\theta_j$ sachant les valeurs passées de $\pi$ et $\sigma_j$ sont déterminés par la méthode du maximum de vraisemblance et puisque les $\theta_j$ sont invariants dans le temps, nous estimons aussi par la méthode du maximum de vraisemblance la valeur de $\pi$ sachant les $\sigma_j$ et les $\theta_j$.

Pour la distance de Kendall (2), nous avons montré dans Fétiveau et al. (2024) que l'estimateur du maximum de vraisemblance $\hat{\theta}$ de $\theta$ est solution de l'équation suivante, voir aussi Fligner et Verducci (1986) :

$$\frac{n\exp(-\hat{\theta})}{1-\exp(-\hat{\theta})} - \sum_{k=1}^{n} \frac{k\exp(-k\hat{\theta})}{1-\exp(-k\hat{\theta})} = \frac{1}{T}\sum_{t=1}^{T} d(\sigma_t, \pi_t), \tag{3}$$

où $t \in \{1, \ldots, T\}$ représente une situation financière des produits de l'entreprise à un instant $t$ et $n$ la taille des permutations. Sachant que la distance de Kendall est invariante à la relabélisation, il n'est pas nécessaire que tous les éléments $\pi_t$ soient les mêmes aux instants $t$. Puisque nos données sont collectées à intervalle de temps régulier, il est important que la permutation $\pi_t$ n'impacte pas les résultats.

Le comportement asymptotique de l'estimateur $\hat{\theta}$ de $\theta$ est donné par le Théorème 1 ci-dessous.

**Théorème 1** *Soit le modèle $(\mathcal{S}_n, \{P_\theta\}_{\theta\in\Theta})$ un modèle régulier où $\mathcal{S}_n$ est l'espace des permutations tel que pour chaque $\theta \in \Theta$, il existe un voisinage $V \subset \Theta$ de $\theta$ pour lequel $\sup_{\alpha\in V} ||\nabla^2 \ln L(.;\alpha)|| \in \mathbb{L}^1(P_\theta)$. Puisque l'estimateur du maximum de vraisemblance $\hat{\theta}$ de $\theta$ est consistant alors*

$$\sqrt{T}\left(\hat{\theta} - \theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta)^{-1}\right) \tag{4}$$

*où*

$$I(\theta) = \frac{n\exp(-\theta)}{(1-\exp(-\theta))^2} - \sum_{k=1}^{n} \frac{k^2\exp(-k\theta)}{(1-\exp(-k\theta))^2}. \tag{5}$$

L'estimateur de la variance asymptotique de $\hat{\theta}$ est donné par

$$I(\hat{\theta})^{-1} = \left(\frac{n\exp(-\hat{\theta})}{(1-\exp(-\hat{\theta}))^2} - \sum_{k=1}^{n} \frac{k^2\exp(-k\hat{\theta})}{(1-\exp(-k\hat{\theta}))^2}\right)^{-1}. \tag{6}$$

Pour chaque $j \in \{1, ..., J\}$, $\sigma_j$ est connu et son paramètre de dispersion $\theta_j$ est estimé par l'estimateur du maximum de vraisemblance $\hat{\theta}_j$. Nous estimons $\pi$ par la méthode du maximum de vraisemblance et nous notons $\hat{\pi}$ son estimateur. En supposant les $\sigma_j$ indépendants, la vraisemblance s'écrit :

$$L(\pi|\sigma_1,\ldots,\sigma_J,\hat{\theta}_1,\ldots,\hat{\theta}_J) = \prod_{j=1}^{J} \frac{\exp(-\hat{\theta}_j\ d(\sigma_j,\pi))}{Z(\hat{\theta}_j)} \tag{7}$$

où $J$ est le nombre de critères. Puisque $Z(\hat{\theta}_j)$ ne dépend pas de $\pi$ et que l'exponentielle est une fonction monotone, l'estimateur du maximum de vraisemblance $\hat{\pi}$ de $\pi$ est :

$$\hat{\pi} = \underset{\pi\in\mathcal{S}_n}{\operatorname{argmax}} \sum_{j=1}^{J} -\hat{\theta}_j\ d(\sigma_j,\pi) \tag{8}$$

où $\mathcal{S}_n$ est l'ensemble de toutes les permutations possibles de n éléments.

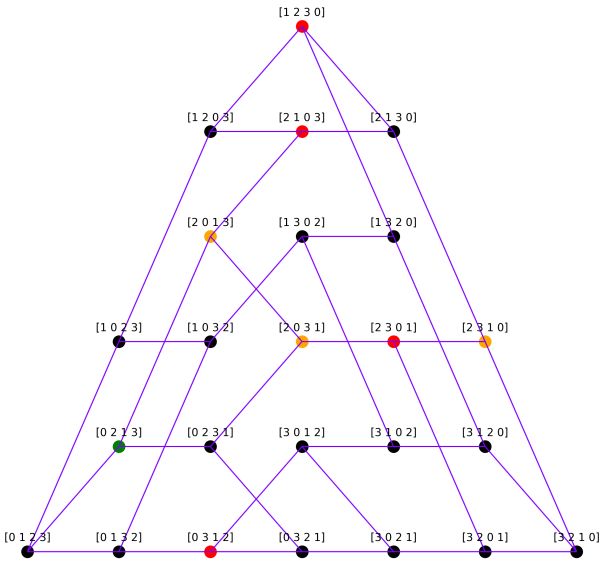



Figure 1: Choix de représentation de trois exemples de consensus en orange pour des critères en rouge et une vérité en vert. Chaque lien de la représentation graphique indique une distance de Kendall de 1 entre deux permutations.

La Figure 1 montre pour des permutations de taille 4, des exemples de consensus $\hat{\pi}$ en orange pour 4 critères, ayant chacun une valeur de $\theta$ fixée arbitrairement à la valeur 1. Les permutations $\sigma_j$ représentant les classements de ces critères sont affichées en rouge. Tous ces consensus n'ont pas la même proximité avec la permutation $\pi$ que l'on cherche à estimer (en vert). La distance de Kendall entre deux permutations peut se déterminer en comptant le nombre de liens minimums pour passer de l'une à l'autre.

# 3 Simulation du modèle de Mallows

Puisque la distance de Kendall est décomposable en une somme finie de termes indépendants, on peut l'utiliser pour simuler des permutations $\sigma$ pour $\pi$ et $\theta$ connus.

En utilisant la distance de Kendall avec le modèle de Mallows ou de Mallows généralisé, Fligner et Verducci (1986) et Fétiveau et al. (2024), il est possible de générer une permutation via le produit des probabilités de déplacer chaque élément de $\pi$ avec les éléments lui succédant. Pour chaque élément de $\pi$, le déplacement d'un élément avec un nombre d'éléments lui succédant est indépendant du mouvement des éléments le précédant. On note $\zeta(\pi^{-1}(i))$ le nombre de déplacements du $i$-ème élément de la permutation $\pi$ avec les éléments le succédant. La probabilité de déplacer un élément d'un nombre $x \leq (n-i)$ d'éléments lui succédant est donnée par

$$\mathbb{P}_\theta\left(\zeta(\pi^{-1}(i)) = x\right) = \frac{\exp(-\theta\ x)}{\sum_{k=0}^{n-i} \exp(-\theta\ k)}. \tag{9}$$

On peut réécrire le dénominateur avec la propriété de la somme d'une suite géométrique.

En utilisant (9), nous représentons dans la Figure 2 un ensemble de 10 000 permutations générées autour de la permutation identité $\pi = \mathrm{Id} = [0\ 1\ 2\ 3\ 4]$ pour $\theta = 0.5$ en considérant un modèle de Mallows à $n = 5$ éléments. La couleur sur ce graphique représente la fréquence d'apparition d'une permutation.

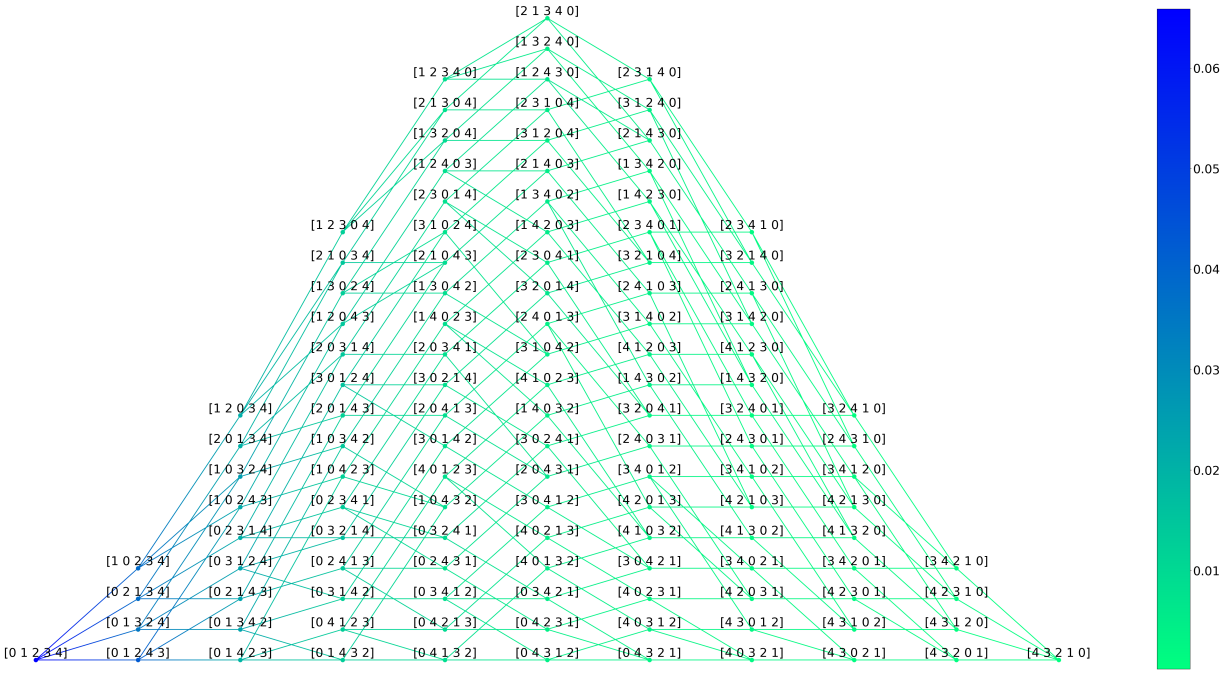



Figure 2: Simulation de 10 000 permutations par le modèle de Mallows. Chaque lien indique une distance de Kendall de 1 entre deux permutations. Plus les permutations sont bleues et plus la fréquence d'apparition est élevée. Plus elles sont vertes et plus la fréquence d'apparition est faible.

La Figure 3 représente une simulation de 10 000 permutations selon un modèle de Mallows généralisé ayant pour vecteur de paramètres $\boldsymbol{\theta} = (0.1, 0.5, 1, 1)$. On observe notamment la forte probabilité de mouvement du premier élément en observant la ligne bleue qui passe sur

les permutations où l'élément 0 se déplace, c'est à dire le premier élément de la permutation initiale $\pi = \mathrm{Id} = [0\ 1\ 2\ 3\ 4]$ se déplace.

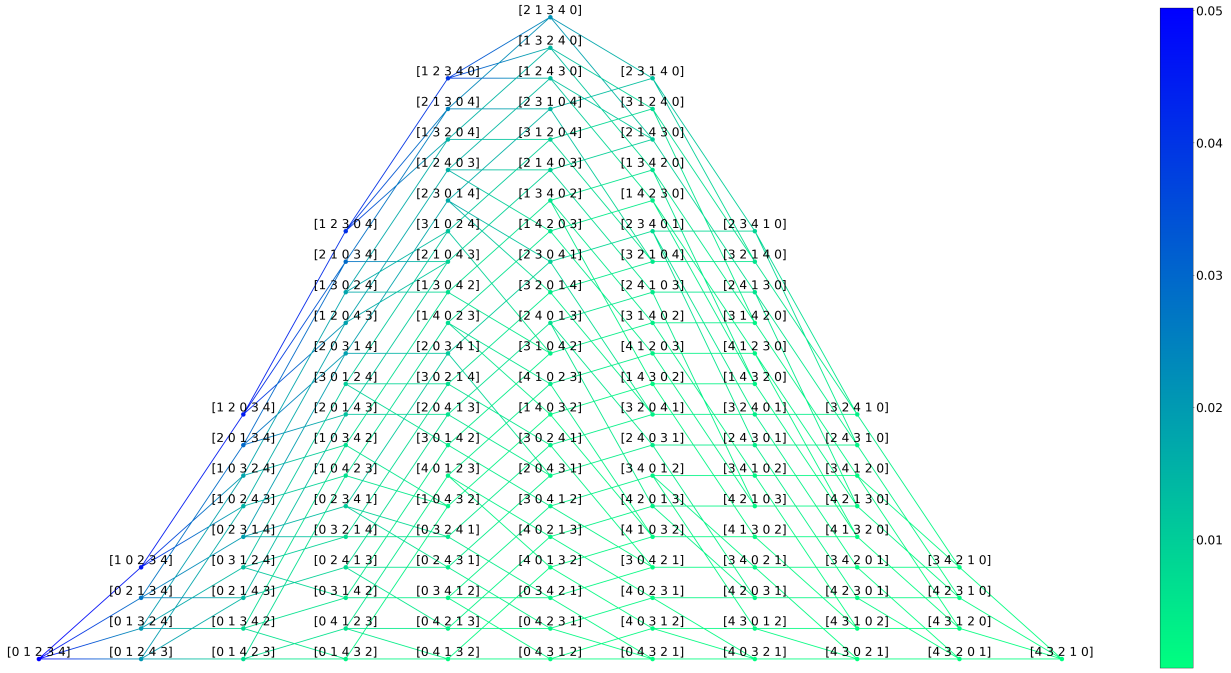



Figure 3: Simulation de 10 000 permutations selon un modèle de Mallows généralisé. Chaque lien indique une distance de Kendall de 1 entre deux permutations. Plus les permutations sont bleues et plus la fréquence d'apparition est élevée. Plus elles sont vertes et plus la fréquence d'apparition est faible.

# 4 Dépendance dans le modèle

Jusqu'ici, nous partions du principe que l'on choisit un ensemble de critères utiles et indépendants. Cependant, rien ne justifie dans les applications que nous sachions quel critère est utile et en l'absence de garantie que l'on connaisse tous les critères utiles, certains critères corrélés peuvent apporter de l'information. On s'intéresse donc à modifier légèrement nos paramètres estimés $\hat{\theta}_j$ de façon à pénaliser deux critères similaires, apportant par conséquent plusieurs fois la même information et attirant le consensus.

Pour améliorer nos estimations $\hat{\theta}$ au sens de la distance minimale entre $\pi$ et $\hat{\pi}$, il semble judicieux d'apprendre les paramètres en minimisant la fonction de coût $\Psi$ donnée par :

$$\Psi(\theta_1, \ldots, \theta_J) = \sum_{t=1}^{T} d(\hat{\pi}_t, \pi_t) \tag{10}$$

où $\hat{\pi}_t$ est estimé en utilisant (8) une fonction de $\theta_1, \ldots, \theta_J$, avec $\sigma_{1t}, \ldots, \sigma_{Jt}$ connus.

L'estimateur $(\hat{\theta_1}, \ldots, \hat{\theta_J})$ maximise la fonction $\Psi$ donnée en (10).

Cependant, cette fonction est une fonction en escalier. Elle reste donc constante en modifiant légèrement $\theta$ jusqu'à ce qu'un changement de consensus ait lieu, moment auquel le coût va changer. On a donc besoin d'aider l'apprentissage à se faire en créant une pente artificielle. Pour chaque instant $t$, on va se baser sur le critère que l'on utilise pour estimer $\hat{\pi}_t$. On sait que $\hat{\pi}_t$ est une permutation maximisant $\sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \hat{\pi}_t)$ donc minimisant la distance pondérée à chaque $\sigma_{jt}$, mais rien ne force $\pi_t$ à maximiser cette somme. Par conséquent, on ajoute à notre fonction de coût $\Psi$ la différence entre $\sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \hat{\pi}_t)$ et $\sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \pi_t)$. Cela a pour conséquence de forcer la distance pondérée de la permutation réelle $\pi_t$ aux $\sigma_{jt}$ à se rapprocher de la distance pondérée de la permutation estimée $\hat{\pi}_t$ aux $\sigma_{jt}$. On obtient donc la fonction de coût

$$\tilde{\Psi}(\theta_1, \ldots, \theta_J) = \sum_{t=1}^{T} \left( \; d(\hat{\pi}_t, \pi_t) \; + \; \left( \sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \hat{\pi}_t) \; - \; \sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \pi_t) \right) \right). \qquad (11)$$

La deuxième partie de ce coût n'étant pas constante, on peut utiliser une descente de gradient pour l'optimiser.

# 5 Étude par simulation

Nous proposons une étude par simulation pour tester les performances de notre approche selon le critère de proximité entre $\pi$ et $\hat{\pi}$.

Nous générons 4 critères utiles avec un modèle de Mallows autour de $\pi$ avec des paramètres $\theta_j$ uniformément aléatoires dans un intervalle fixé arbitrairement et excluant 0 (voir Section 3). Nous générons 1 critère inutile, une permutation simulée par une loi uniforme sur l'ensemble de l'espace des permutations $\mathcal{S}_n$, et 3 critères dérivés de 2 des critères utiles. Les critères dérivés sont déterminés à partir de variables sous-jacentes aux critères utiles. Celles-ci sont générées selon des lois log-normales de paramètres aléatoires dans des intervalles fixés arbitrairement et triées selon l'ordre indiqué par la permutation du critère. Nos critères dérivés sont donc la permutation associée au résultat d'une fonction des variables sous-jacentes des critères utiles.

On simule donc 360 ensembles aléatoires de paramètres. Chaque ensemble de paramètres, représentant les $\theta$ réels et les paramètres des lois log-normales, est utilisé pour générer 500 jeux de données de 5 éléments. On a donc au total 180 000 jeux de données de 5 éléments, tous avec les mêmes structures de critères mais des paramètres différents.

On va donc comparer la distance entre $\pi$ et $\hat{\pi}$ pour chaque ensemble de $\theta$ estimé avec la formule (3) en ne prenant que les critères utiles puis avec tous les critères. Puis cette même distance avec des $\theta$ ré-estimés en utilisant la fonction de coût (11) où les $\theta$ sont initialisés par l'estimateur du maximum de vraisemblance (3) puis aléatoirement entre 0 et 1. On compare donc 4 possibilités d'estimation des $\theta$, respectivement notées 'EMV variables utiles', 'EMV', 'EMV+Optim' et 'Optim'.

Pour chaque ensemble de paramètres, on va s'entraîner sur 300 jeux de données et tester sur 200.

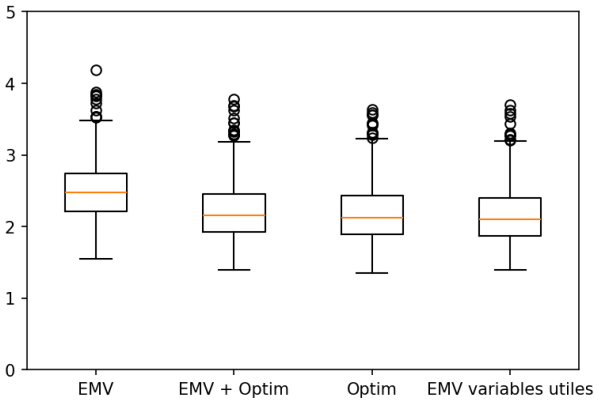



Figure 4: Boxplot de l'écart entre $\pi$ et $\hat{\pi}$ pour différentes estimations de $\theta$ dans les jeux d'entraînements.

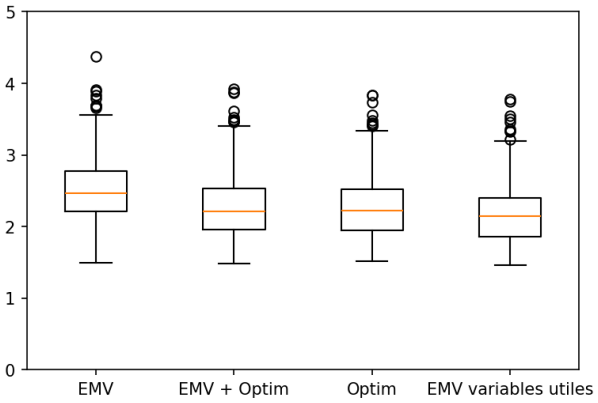



Figure 5: Boxplot de l'écart entre $\pi$ et $\hat{\pi}$ pour différentes estimations de $\theta$ dans les jeux de tests.

Dans la Figure 4, nous observons la distribution de la distance entre $\pi$ et $\hat{\pi}$ pour l'ensemble des jeux de paramètres. Cette distance est théoriquement comprise entre 0 et 10 mais 10 étant le classement inverse de 0, le pire pour nous se situe à une distance de 5 que l'on obtiendrait en tirant $\hat{\pi}$ aléatoirement. Chaque jeu de paramètre est composé de 300 jeux de données sur lesquels nous avons estimé un $\theta$ global puis nous avons calculé la distance moyenne entre $\pi$ et $\hat{\pi}$. Nous obtenons donc 360 distances moyennes pour chaque évaluation de $\theta$. La Figure 5 représente quant à elle les écarts calculés sur les 200 jeux de données de tests des 360 ensembles de paramètres. Nous observons une nette diminution de la distance entre $\pi$ et $\hat{\pi}$ lors de l'optimisation avec notre fonction de coût (11) que ce soit en initialisant à partir des estimateurs du maximum de vraisemblance ou aléatoirement. Dans l'entraînement, alors que

nous utilisons toutes les variables, qu'elles soient utiles, dérivées ou inutiles, nous arrivons à approcher les résultats de l'utilisation exclusive des variables utiles. Pour l'entraînement, nous avons en moyenne une distance de 2.52 pour l'EMV, 2.23 pour l'optimisation à partir de l'EMV, 2.21 pour l'optimisation à initialisation aléatoire et 2.17 pour l'EMV des variables utiles. Nous éliminons donc 83% de la distance ajoutée avec les variables inutiles et dérivées en optimisant à partir de l'EMV et 89% de cette distance ajoutée en optimisant à partir d'une initialisation aléatoire. Pour ce qui est du test, nous avons en moyenne une distance de 2.53 pour l'EMV, 2.29 pour l'optimisation à partir de l'EMV, 2.28 pour l'optimisation à initialisation aléatoire et 2.18 pour l'EMV des variables utiles. Nous éliminons donc 69% de la distance ajoutée avec les variables inutiles et dérivées en optimisant à partir de l'EMV et 71% de cette distance ajoutée en optimisant à partir d'une initialisation aléatoire.

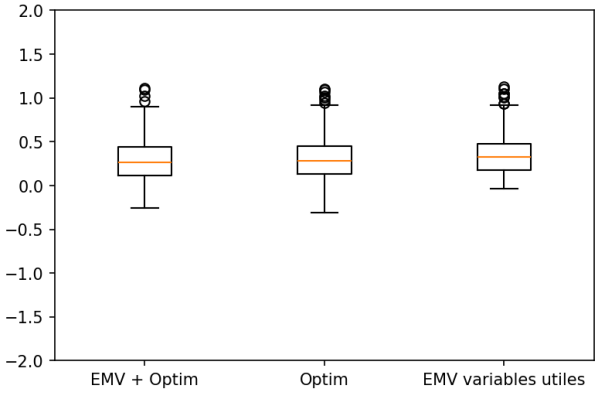



Figure 6: Boxplot de l'écart à l'EMV en terme de distance entre $\pi$ et $\hat{\pi}$ pour différentes estimations de $\theta$ dans les jeux d'entraînement.

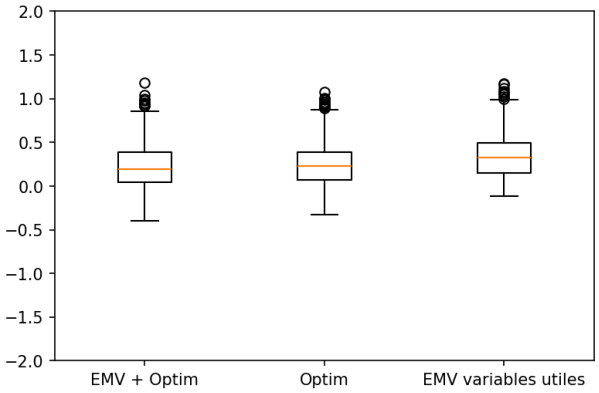



Figure 7: Boxplot de l'écart à l'EMV en terme de distance entre $\pi$ et $\hat{\pi}$ pour différentes estimations de $\theta$ dans les jeux de tests.

Les Figures 6 et 7 montrent la différence de distance de chaque estimation avec l'EMV. Même si globalement ces estimations sont meilleures, il leur arrive d'être inférieures à l'EMV.

Sur les 360 ensembles de paramètres nous avons 40 ensembles de paramètres avec la méthode d'optimisation à partir de l'EMV, 19 ensembles de paramètres avec la méthode d'optimisation à partir de $\theta$ aléatoires et 3 ensembles de paramètres à partir de l'EMV des variables utiles qui sont moins bons qu'avec l'estimation par l'EMV pour les jeux d'entraînement. Pour le test, nous avons 63 ensemble de paramètres avec la méthode d'optimisation à partir de l'EMV, 59 ensembles de paramètres avec la méthode d'optimisation à partir de $\theta$ aléatoires et 15 ensembles de paramètres à partir de l'EMV des variables utiles qui sont moins bons que l'EMV.

# 6 Conclusion

Nous avons montré que l'optimisation de la fonction (11) permet d'éliminer, sans savoir quelles sont les variables utiles, une bonne partie de l'erreur ajoutée par les variables dérivées et inutiles. Cependant, cela nécessite un temps de calcul plus important lors de l'apprentissage que pour l'estimateur du maximum de vraisemblance.

# Bibliographie


Diaconis, P. (1988), Group representations in probability and statistics, *Lecture notes-monograph series*, 11, p. 112.

Fétiveau, A., Durrieu, G., Frénod, E., Meledo, C. H. and Prat, B. (2024), Permutations based model for business performance, *Discrete and Continuous Dynamical Systems Series S*, in revision.

Fligner, M. A., and Verducci, J. S. (1986), Distance based ranking models, *Journal of the Royal Statistical Society: Series B*, 48(3), pp. 359-369.

Keeney, R. L., and Raiffa, H. (1976), Decisions with multiple objectives: Preferences and value tradeoffs, John Willey and Sons, New York.

Kendall, M. G. (1938), A New Measure of Rank Correlation, *Biometrika*, 30, pp. 81-93

Mallows, C. L. (1957), Non-null ranking models, *Biometrika*, 44(1/2), pp. 114-130.

Mareschal, B., Brans, J. P., and Vincke, P. (1984), PROMETHEE: A new family of outranking methods in multicriteria analysis, *In: Operational Research*, Elsevier Science Publishers B.V., pp. 408-421.

Roy, B. (1968), Classement et choix en présence de points de vue multiples, *Revue française d'informatique et de recherche opérationnelle*, 2(8), pp. 57-75.

# Modélisation statistique pour l'identification et la quantification de manipulations comptables

Marie Chavent[1], Véronique Darmendrail[2], Delphine Feral[1], Hadrien Lorenzo[3], Frédéric Pourtier[2], Jérôme Saracco[1]

[1] *Univ. Bordeaux, CNRS, INRIA, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France*
marie.chavent@inria.fr, delphine.feral@u-bordeaux.fr, jerome.saracco@inria.fr

[2] *Univ. Bordeaux, IRGO, UR 4190, F-33000 Bordeaux, France*
veronique.darmendrail@u-bordeaux.fr, frederic.pourtier@u-bordeaux.fr

[3] *Univ. Aix-Marseille, I2M, UMR 7373, Marseille, France*
hadrien.lorenzo@univ-amu.fr

**Résumé.** Cette communication présente une nouvelle méthodologie statistique pour détecter la manipulation comptable autour d'un seuil psychologique (ici, le bénéfice nul d'une entreprise). Un algorithme de type EM est proposé pour estimer les paramètres du modèle de manipulation comptable. Le bon comportement numérique de la méthodologie est illustré sur des données simulées proches des données réelles.

**Mots-clés.** Statistique appliquée, algorithme EM ; estimation de la densité ; modèle de mélange ; distribution gaussienne tronquée ; manipulation comptable.

**Abstract.** This communication presents a new statistical methodology to detect earnings management associated with the zero earnings threshold. An EM-type algorithm is proposed to estimate the underlying parameters of the considered earnings management model. The good numerical behavior of the methodology is illustrated on simulated data close to real data.

**Keywords.** Applied Statistics, EM algorithm; density estimation; Mixture model; Truncated Gaussian distribution; Earnings management.

# 1 Motivation

La pratique de la manipulation comptable par les entreprises est reconnue depuis longtemps par les chercheurs. La littérature comptable a montré qu'elles sont enclines à gérer leurs bénéfices et leurs pertes comptables de manière à dépasser des seuils spécifiques tels que le bénéfice nul, le bénéfice de l'année précédente ou les prévisions des analystes, voir par exemple Burgstahler & Chuk (2017), ou encore Byzalov & Basu (2019). L'existence de manipulations comptables compromet l'intégrité de l'information financière. Les régulateurs ou les investisseurs s'intéressent donc à la détection de la manipulation comptable, à sa fréquence et à son ampleur. Dans cette communication, une nouvelle méthodologie statistique pour détecter la manipulation comptable associée au seuil de bénéfice nul est présentée. Le modèle statistique de manipulation comptable est présenté à la Section 2. La Section 3 donne une brève description de la procédure d'estimation des paramètres du modèle proposé. Cette dernière est illustrée numériquement à la Section 4 sur des données simulées très similaires aux données réelles.

# 2 Présentation du modèle statistique de manipulation comptable

Pour déterminer la fréquence et l'ampleur de la manipulation comptable (autour du seuil psychologique correspondant au bénéfice nul) dans un échantillon d'entreprises, la distribution des bénéfices doit être modélisée. Désignons par $X_i$ (resp. $Y_i$) le bénéfice réel (resp. observé) de l'entreprise $i$, $i = 1, \ldots, N$. Notons que les revenus réels sont supposés être indépendants et identiquement distribués et suivre un **mélange de deux distributions gaussiennes** de densité

$$\boldsymbol{\pi\varphi_A(x) + (1-\pi)\varphi_B(x)}, \tag{1}$$

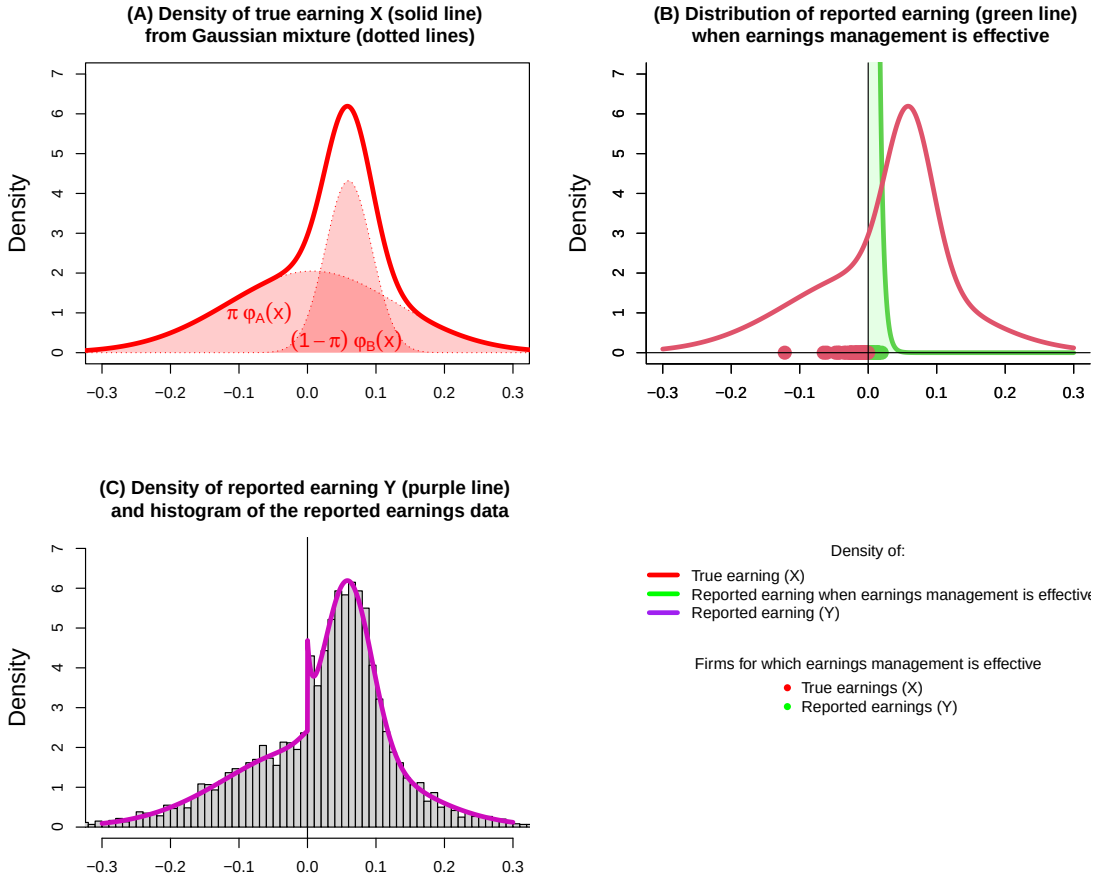
où $\pi \in ]0,1[$ et $\varphi_A$ (resp. $\varphi_B$) est la densité de la loi normale d'espérance $\mu_A$ (resp. $\mu_B$) et de variance $\sigma_A^2$ (resp. $\sigma_B^2$). Le choix d'un mélange gaussien permet de modéliser une distribution avec des queues de poids différents, comme cela est souvent observé dans les données comptables que l'on va considérer.

Lorsqu'une entreprise a une valeur de bénéfice réel inférieure au seuil zéro, elle peut s'engager ou pas dans une manipulation comptable et le bénéfice déclaré sera alors supérieur au seuil zéro. Lorsque cette manipulation comptable est effective pour une entreprise $i$, le revenu déclaré $Y_i$ est supposé suivre la **loi exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda$**, sinon $Y_i = X_i$ le revenu réel.

Afin de modéliser la transition vers la manipulation comptable (des bénéfices) pour une entreprise $i$, la variable aléatoire $T_i$ est introduite : lorsque $X_i$ est inférieur à la valeur seuil de zéro, $T_i$ sachant $X_i = x$ suit une distribution de Bernoulli de paramètre $\tau(x)$. Cette fonction $\tau(.)$ définie pour $x < 0$ est croissante, de manière à ce que la probabilité de manipulation soit d'autant plus forte que le bénéfice de l'entreprise est proche du seuil psychologique zéro.

L'idée sous-jacente est de se concentrer uniquement sur le sous-échantillon d'entreprises

Figure 1: Les densités sous-jacentes à la modélisation de la manipulation comptable.



associées aux $Y_i \geq 0$, c'est à dire celles ayant déclaré un bénéfice positif. Le travail ci-après est donc implicitement développé conditionnellement à $Y \geq 0$. Les bénéfices déclarés correspondants de ces entreprises peuvent alors être considérés comme un mélange entre les bénéfices "réels" et les bénéfices "manipulés", qui peuvent être modélisés comme suit :

$$\forall \boldsymbol{y \geq 0}, \quad \boldsymbol{f(y) = q f_1(y) + (1-q) f_2(y)}, \tag{2}$$

où $q \in ]0,1[$, $f_1$ est la densité de la loi $\mathcal{E}(\lambda)$, et $f_2$ est la densité du mélange:

$$\forall y \geq 0, \quad f_2(y) = \tilde{\pi}\varphi_A^{(+)}(y) + (1-\tilde{\pi})\varphi_B^{(+)}(y), \tag{3}$$

avec $\varphi_A^{(+)}$ (resp. $\varphi_B^{(+)}$) la densité de la loi normale tronquée (à zéro) d'espérance $\mu_A$ (resp. $\mu_B$) et de variance $\sigma_A^2$ (resp. $\sigma_B^2$). Le vecteur des paramètres sous-jacents du modèle simplifié[1] est donc le suivant :

$$\theta = \left(q, \lambda, \tilde{\pi}, \mu_A, \sigma_A^2, \mu_B, \sigma_B^2\right). \tag{4}$$

# 3 Procédure d'estimation des paramètres

Étant donné l'échantillon observé des revenus (positifs) déclarés $\mathcal{D}_n = \{y_1, \ldots, y_n\}$, l'objectif est d'estimer $\theta$ par maximum de vraisemblance

$$\hat{\theta} = \left(\hat{q}, \hat{\lambda}, \hat{\tilde{\pi}}, \hat{\mu}_A, \hat{\sigma}_A^2, \hat{\mu}_B, \hat{\sigma}_B^2\right) := \arg\max_{\theta} \ \ell(\theta; \mathcal{D}_n) \tag{5}$$

où $\ell(\theta, \mathcal{D}_n) = \prod_{i=1}^n f(y_i)$.

A cette fin, un algorithme de type EM (Expectation–Maximization) (voir Dempster et al., 1977) a été mis au point. Cet algorithme repose sur la vraisemblance complétée par deux variables latentes dichotomiques : $\tilde{T}_i$ (resp. $\tilde{Z}_i$) qui suit une distribution de Bernoulli de paramètre $q$ (resp. $\tilde{\pi}$). La variable latente $\tilde{T}_i$ indique si l'entreprise $i$ a manipulé ses bénéfices ou non, tandis que la variable $\tilde{Z}_i$ est utilisée pour gérer le mélange des deux distributions gaussiennes tronquées. La densité "complétée", élément de base pour mettre en œuvre l'algorithme EM, est donnée par :

$$\left(q f_1(y)\right)^t (1-q)^{1-t} \left(\tilde{\pi}\varphi_A^{(+)}(y)\right)^{(1-t)z} \left((1-\tilde{\pi})\varphi_B^{(+)}(y)\right)^{(1-t)(1-z)}. \tag{6}$$

Plus de détails techniques sont disponibles dans Chavent et al. (2023).

[1] dans le sens où l'on ne se focalise que sur la sous-échantillon d'entreprises ayant déclaré un bénéfice positif

# 4 Illustration numérique

Une rapide description des données simulées est fournie à la Section 4.1. Les résultats obtenus sont commentés brièvement à la Section 4.2.

## 4.1 Paramètres des simulations

Deux scénarios ont été envisagés. Deux échantillons de taille $n = 2000$ ont été générés à partir du modèle (2) pour les deux paramètres $\theta$ décrits dans la Table 1.

Table 1: Paramètres utilisés dans les simulations.

| | $q$ | $\lambda$ | $\tilde{\pi}$ | $\mu_A$ | $\sigma_A$ | $\mu_B$ | $\sigma_B$ |
|---|---|---|---|---|---|---|---|
| Scénario 1 | 0.1 | 12 | 0.3 | 1 | 1 | 4 | 0.5 |
| Scénario 2 | 0.05 | 185.3 | 0.632 | 0.007 | 0.123 | 0.06 | 0.034 |

**Commentaires rapides sur les deux scénarios.**

- Dans le premier scénario, il est relativement facile d'identifier et d'estimer toutes les composantes de $\theta$.
- Le second scénario est plus complexe, mais correspond à des situations plus réalistes dans le contexte de la manipulation comptable. Il est fortement inspiré du premier jeu de données réelles qui a été utilisé par Chen et al. (2010).

## 4.2 Résultats

Les résultats numériques obtenus sont respectivement présentés dans les figures 2 et 3. Afin d'avoir une visibilité sur la variabilité des estimateurs, $B = 100$ échantillons Bootstrap ont été générés, et les paramètres associés ont ensuite été estimés. La variabilité de la densité estimée est fournie via l'intervalle de confiance Bootstrap à 90% (en rouge), ainsi que celles de l'estimation des composantes de $\theta$ via les boxplots correspondants.

On observe clairement que la procédure d'estimation permet de retrouver correctement les composantes du paramètre $\theta$ et donc la vraie densité $f$ des $Y_i$, aussi bien avec les données issues du scénario 1 (voir Figure 2) qu'avec les données issues du scénario 2 (voir Figure 3).

Figure 2: Résultats de la simulation dans le cadre du scénario 1 avec la vraie densité $f$ (en bleu) et la densité estimée (en marron), ainsi que les valeurs réelles et estimées de $\theta$, avec la variabilité bootstrap (en orange pâle).

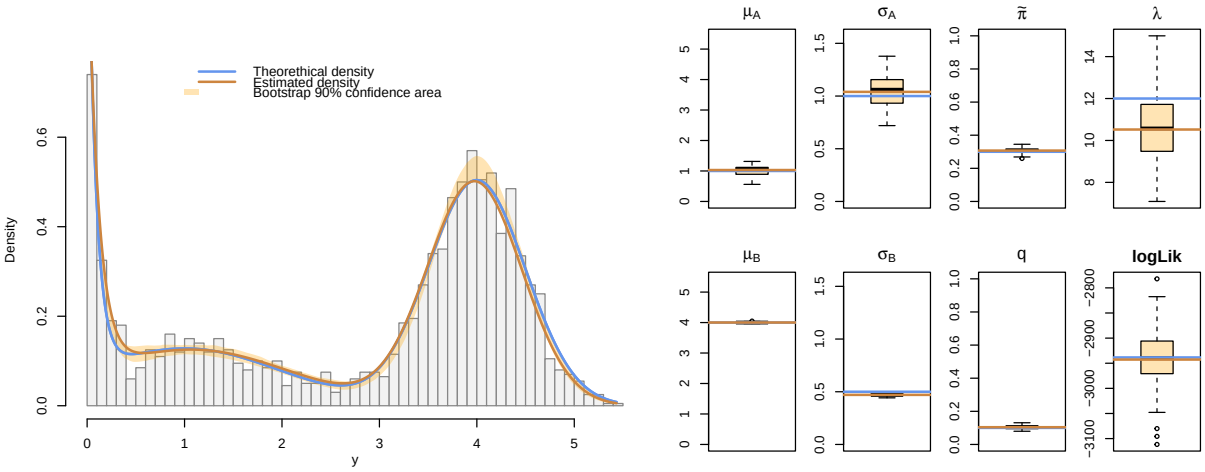
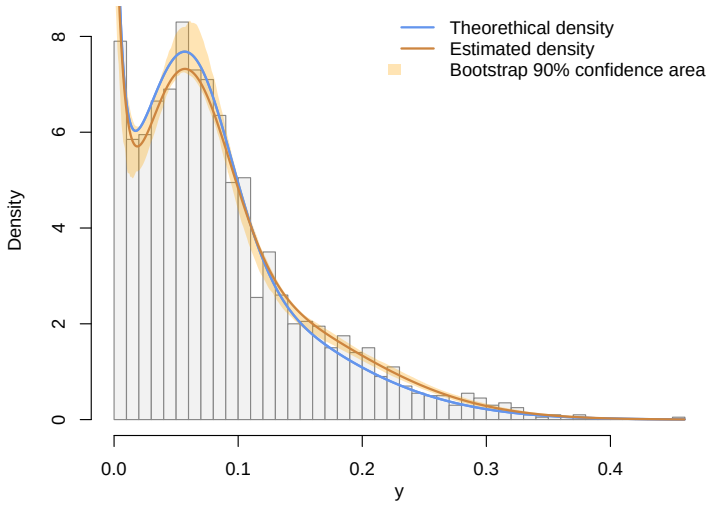



Figure 3: Résultats de la simulation dans le cadre du scénario 2 avec la vraie $f$ densité (en bleu) et la densité estimée (en marron) et la variabilité bootstrap (en orange pâle).



# 5 Observations finales

- Dans le cadre de ce résumé, le bon comportement numérique de la procédure d'estimation n'a été évalué que sur deux jeux de données (associés à deux scénarios différents) à la Section 4. Des simulations numériques de plus grande ampleur seront présentées lors de la communication orale.

  Une application sur données réelles servira également d'illustration.

- À partir du paramètre estimé $\hat{\theta}$ du modèle simplifié, il est possible d'obtenir une estimation des paramètres du modèle statistique initial[2] et de retrouver la fréquence et l'ampleur de la manipulation comptable sur l'ensemble de la population des entreprises.

- En se basant sur la variabilité Bootstrap, il sera également possible de comparer des modèles basés sur deux populations, comme des pays différents, des sous-périodes différentes, ou différentes catégories d'entreprises.

  Ce point sera illustré sur des données réelles avec une comparaison entre deux pays et une comparaison avec deux méthodes comptables.

- De plus, il existe d'autres domaines d'application pour ces méthodes de détection et de quantification de manipulation de données autour d'un seuil psychologique. Dans différents domaines (autres que celui de la comptabilité), un décideur a parfois l'opportunité et l'incitation de passer d'un niveau juste inférieur ou à un niveau juste supérieur à un point de référence, grâce à de la manipulation de données.

- Un package R est actuellement en cours de développement.

# Bibliographie


Burgstahler, D. and Chuk, E. (2017). What have we learned about earnings management? Integrating discontinuity evidence. *Contemp. Account. Res.*, 34(2), 726749.

Byzalov, D. and Basu, S. (2019). Modeling the determinants of meet-or-just-beat behavior in distribution discontinuity tests. *Journal of Accounting and Economics*, 68(23), 101266.

Chavent, M., Darmendrail, V., Feral, D., Lorenzo, H. , Pourtier, F. and Saracco, J. (2023) A new statistical methodology to detect earnings management. *Proceedings of the 37th International Workshopon Statistical Modelling (IWSM 2023)*, July 17-21, 2023-Dortmund, Germany, 394 – 399.

Chen, S.K., Lin, B-X., Wang, Y. and Wu, L. (2010). The frequency and magnitude of earnings managements: Time-series and multi-threshold comparisons. *International Review of Economics and Finance*, **19**, 671 – 685.

Dempster, A.P., Laird, N.M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc., Ser. B*, **39**, 1 – 38.


[2] c'est-à-dire du modèle qui ne se limite pas uniquement aux bénéfices positifs.

# Conférence Lucien Le Cam

# Local Asymptotic Optimality in Empirical Bayes, Bias Correction and Benign Overfitting

Cun-Hui Zhang

Rutgers University

We present examples to demonstrate the great power and broad impact of Le Cam's celebrated local asymptotic theory. In empirical Bayes, an extension of the Hájek–Le Cam convolution theorem applies to the estimation of functionals of both the data and parameters motivated by risk evaluation, species and network problems. In semi-low-dimensional models, Le Cam's one-step estimator leds to de-biased inference, and local asymptotic normality provides the direction of efficient projection in bias correction. Akin to local asymptotic minimaxity, the adaptive minimaxity of empirical Bayes estimators is discussed, including an application of general maximum likelihood empirical Bayes to linear regression. Finally, in an explanation of the double descent phenomenon in machine learning, Le Cam's one step method provides general local optimality of interpolation estimators in high-dimension.

# Session spéciale Statistique & Santé

# Modéliser et communiquer les évidences et les sources d'incertitudes pour améliorer la replicabilité et la crédibilité de la recherche biomédicale

Hoffmann Sabine*[1]

[1]Ludwig-Maximilians-Universität München – Allemagne

**Résumé**

La communication des évidences et des incertitudes en science est notoirement difficile. Si la question de l'interprétabilité a récemment fait l'objet d'une attention particulière dans la littérature de machine learning, cette question reste également importante lorsqu'il s'agit de quantités statistiques classiques. Les erreurs d'interprétation des p-valeurs et des estimations des effets, incluant les odds ratios et les hazard ratios, sont si fréquentes que les chercheurs appliqués sont parfois surpris lorsqu'ils entendent la définition correcte de ces quantités. En outre, les études métascientifiques menées ces dernières années ont conduit à une prise de conscience croissante de la multiplicité des stratégies possibles dans l'analyse de données empiriques. Lorsque cette multiplicité est combinée à une publication séléctive des analyses et des résultats, les méthodes statistiques classiques qui ignorent cette multiplicité transmettent un niveau de certitude disproportionné et peuvent conduire à des résultats trop confiants. Dans la recherche biomédicale, la situation est encore compliquée par le fait que les résultats scientifiques sont souvent le fruit de différents types d'études qui peuvent être soumis à des sources d'incertitude très différentes. Ces sources d'incertitude peuvent être encore exacerbées par la disponibilité croissante de données collectées de manière routinière dans lesquelles les mécanismes de génération des données et les procédures de mesure sont peu connus et peu contrôlés. Cette présentation donnera un aperçu des défis actuels en matière de communication des évidences et des incertitudes dans la recherche biomédicale et des idées pour surmonter ces défis afin d'améliorer la replicabilité et la crédibilité des résultats scientifiques. Seront abordés entre autres la modélisation hiérarchique Bayésienne des incertitudes, des méthodes de reporting pour améliorer l'interprétabilité et la comparabilité des résultats de différents modèles, ainsi que la pre-registration d'études observationnelles. En fin de compte, l'objectif sera de communiquer les évidences scientifiques de sorte que le grand public, les praticiens et les décideurs politiques pourront peser les coûts, les avantages et les incertitudes dans leur prise de décision.

**Mots-Clés:** Incertitude, statistique bayésienne, biostatistique, interprétabilité

*Intervenant

# Prédiction dynamique d'événements à partir de multiples marqueurs longitudinaux par « model averaging »

Hélène Jacqmin-Gadda[1], Taban Baghfalaki[1], Reza Hashemi[2]

[1]Inserm, Research Center U1219, Univ. Bordeaux, ISPED, F33076 Bordeaux, France
[2] Department of Statistics, Razi University, Kermanshah, Iran

**Résumé :**

Au cours des dernières années, de nombreux travaux en biostatistique ont porté sur le développement d'outils de prédiction dynamique d'événements de santé à partir de mesures répétées de marqueurs ou facteurs de risque. Ces outils constituent une contribution essentielle au développement de la médecine personnalisée car ils permettent de prédire le risque individuel de survenue d'un événement dans une fenêtre de temps à partir de l'ensemble des informations collectées jusqu'au temps courant. Les modèles conjoints pour un ou plusieurs temps d'événements (potentiellement compétitifs) et les données répétées de marqueurs longitudinaux est la méthode privilégiée dans ce domaine. Ces modèles combinent donc des modèles mixtes et des modèles de survie liés soient par des effets aléatoires partagés soit par des classes latentes. Bien que l'estimation fréquentiste ou Bayésienne de ces modèles soient dorénavant possibles grâce à différents logiciels, l'estimation de modèles conjoints incluant de nombreux marqueurs longitudinaux reste un défi majeur. Quelle que soit l'approche, l'estimation conjointe devient impossible lorsque le nombre de marqueurs est trop grand en raison du nombre trop élevé d'effets aléatoires et de paramètres à estimer et à l'imprécision du calcul numérique des intégrales de grandes dimensions.
Dans ce travail, nous proposons d'estimer les prédictions dynamiques individuelles basées sur les mesures répétées de multiples marqueurs par une moyenne pondérée des prédictions estimées à partir de modèles conjoints ne comportant chacun qu'un marqueur. Les poids peuvent dépendre du temps de prédiction. Ils sont estimés en minimisant le score de Brier dépendant du temps. Bien que le temps de calcul global puisse être long, cette approche est toujours réalisable (et aisément parallélisable), même lorsque le nombre de prédicteurs longitudinaux est très élevé. Ses avantages et limites sont évalués par divers scénarios de simulations et comparer aux prédictions du modèle multi-marqueurs dans les scénarios où il est estimable. Cette méthode est utilisée pour prédire le risque de décès dans la cohorte de personnes âgées 3C à partir de 17 prédicteurs longitudinaux et ses capacités prédictives sont comparées à celles de plusieurs approches alternatives.

**Abstract :**

In recent years, much work in biostatistics has focused on the development of dynamic prediction tools for health events, based on repeated measurements of markers or risk factors. These tools make an essential contribution to the development of personalized medicine, as they enable us to predict the individual risk of an event occurring within a time window, based on all the information collected up to the current time. Joint models for one or more (potentially competing) event times and repeated measures of longitudinal markers are the preferred method in this field. These models thus combine mixed models and survival models linked either by shared random effects or by latent classes. Although frequentist or Bayesian estimation of these models is now possible thanks to various open-access packages, estimation of joint models including numerous longitudinal markers remains a major challenge. Whatever the approach, joint estimation becomes impossible when the number of markers is too large, due to the excessive number of random effects and parameters to be estimated, and to the imprecision of the numerical calculation of high-dimensional integrals.

In this work, we propose to estimate individual dynamic predictions based on repeated measurements of multiple markers by a weighted average of predictions estimated from joint models each comprising only one marker. Weights may depend on prediction time. They are estimated by minimizing the time-dependent Brier score. Although the overall computation time can be long, this approach is still feasible (and easily parallelizable), even when the number of longitudinal predictors is very high. Its advantages and limitations are evaluated in various simulation scenarios and compared with the predictions of the multi-marker model in those scenarios where it may be estimated. This method is used to predict the risk of death in the 3C elderly cohort from 17 longitudinal predictors, and its predictive abilities are compared with those of several alternative approaches

# Detecting genomic alteration in genomic profiles: the infinite population case

Marie-Pierre Etienne*[1], Stéphane Robin , Gabriel Lang , Pierre Vallois , and Laurent Decreusefond

[1]Institut de Recherche Mathématique de Rennes (IRMAR, Agrocampus Ouest) – Agrocampus Ouest – France

**Résumé**

Two states Markov Jump process can be used to model alterations in genomic profiles along a chromosom (0 for normal state and 1 for alteration) in a normal cell. Detecting recurrent alterations among a set of patients based on genomic profiles help to identify genomic regions and potentially genes involved in the disease process. This may be formalized within a statistical test procedure and require to characterize the lengths of the excursions above a given threshold for the process of the cumulated profiles. This work has been done when the size of the cohort is small. When the size of the population increases, we prove that the cumulated process tends to an Ornstein Uhlenbeck (OU) process and we have a bound for the rate of convergence. We prove that this rate of convergence also holds for the convergence of the longest excursion.

**Mots-Clés:** Ornstein Uhlenbeck, excursion, Geomic Alteration

*Intervenant

# Session ENBIS et groupe Fiabilité et Incertitudes : Application industrielle du deep learning

# Physics-informed machine learning et prévision

Nathan Doumèche*[1,2], Gérard Biau , Claire Boyer , and Yannig Goude

[1]Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Universite, Centre National de la Recherche Scientifique, Université Paris Cité, Sorbonne Universite : UMR$_8$001, $CentreNationaldelaRechercheScientifique : UMR_8001, UniversitéParisCité : UMR_8001 - -France$

[2]EDF Labs – EDF Recherche et Développement – France

**Résumé**

L'apprentissage profond avec a priori physique consiste à entraîner des réseaux de neurones en utilisant un risque empirique pénalisé par un système d'équations différentielles (EDP). Ces modèles permettent d'allier la performance des réseaux de neurones à l'interprétabilité apportée par la modélisation physique. Leurs bonnes performances pratiques se sont illustrées dans le cadre de la de la modélisation hybride, combinant un modèle physique imparfait avec des observations bruitées, notamment pour les prévisions climatiques. Cependant, leurs propriétés théoriques restent encore à établir. Ici, nous montrons que l'entraînement classique -et massivement adopté- de ces réseaux peut souffrir d'un surapprentissage systématique. Nous considérons alors l'ajout d'une régularisation de type ridge. Nous montrons alors que l'estimateur ainsi construit est consistant, avec une vitesse de convergence potentiellement accélérée grâce à la contrainte EDP, et qu'il vérifie de surcroît le modèle physique, à l'erreur de modélisation près.

**Mots-Clés:** Physics informed machine learning, Interpétabilité, Vitesse de convergence, Prévision

*Intervenant

# Le Deep Learning pour l'estimation de la distribution en taille de particules de $TiO_2$ à partir d'images en microscopie électronique à balayage

Loïc Coquelin [1] & Paul Monchot [2] & Nicolas Fischer[3] & Nicolas Feltin[4] & Alexandra Delvallée[5]

[1] *Laboratoire National de Métrologie et d'Essais, France, Loic.Coquelin@lne.fr*
[2] *Laboratoire National de Métrologie et d'Essais, France, Paul.Monchot@lne.fr*
[3] *Laboratoire National de Métrologie et d'Essais, France, Nicolas.Fischer@lne.fr*
[4] *Laboratoire National de Métrologie et d'Essais, France, Nicolas.Feltin@lne.fr*
[5] *Laboratoire National de Métrologie et d'Essais, France, Alexandra.Delvallee@lne.fr*

**Résumé.** Afin de détecter et caractériser les particules de $TiO_2$ présentes dans les produits alimentaires et cosmétiques (Weir *et al.* (2012), Hwang *et al.* (2019)), le microscope électronique à balayage (MEB) est généralement l'instrument privilégié avec des images haute résolution permettant de mesurer la taille, la forme ou l' état d'agrégation des particules. Pour comprendre les propriétés des matériaux à l'échelle nanométrique, il s'agit d'estimer précisément la distribution en taille des particules et pour se faire, il faut détecter et segmenter chaque particule individuelle dans l'image et estimer leur état d'agrégation. Nous proposons d'utiliser le Mask-RCNN développé par He *et al.* (2017) pour la tâche de segmentation automatique des particules avec un entraînement couplant les techniques de transfert d'apprentissage et d'augmentation de données compte-tenu du nombre limité d'images annotées (courant dans l'industrie). Pour estimer le statut d'agrégation de chaque particule, nous utilisons un réseau de neurone convolutionnel (VGG16). Cette chaîne de traitement automatisée permet une bonne estimation de la distribution en taille des particules de $TiO_2$ avec 96 % des mesures sur le jeu de test présentant moins de 5 % d'erreur sur l'estimation du diamètre de surface équivalent.

**Mots-clés.** Microscopie Électronique à Balayage, Particules de $TiO_2$, Réseaux de neurones, Segmentation d'instances, Classification ...

**Abstract.** In order to detect and characterize $TiO_2$ particles present in food and cosmetic products (Weir *et al.* (2012), Hwang *et al.* (2019)), the scanning electron microscope (SEM) is generally the preferred instrument with high-resolution images enabling measurement of particle size, shape or aggregation state. To understand the properties of nanoscale materials, we need to accurately estimate the particle size distribution. To do this, we need to detect and segment each individual particle in the image, and estimate their state of aggregation. We propose to use the Mask-RCNN developed by He *et al.* (2017) for the automatic particle segmentation task with training coupling transfer learning and data augmentation techniques given the limited number of annotated images (common in the industry). To estimate the aggregation status of each particle, we use a convolutional neural network (VGG16). This automated pipeline provides a good estimate of the size distribution of $TiO_2$ particles, with 96% of measurements on the test set showing less than 5% error in equivalent surface diameter estimation.

**Keywords.** Scanning Electron Microscopy, $TiO_2$ particles, Neural Networks, Instance Segmentation, Classification ...

# 1 Introduction

A partir d'images acquises au microscope électronique à balayage (MEB), comme l'exemple présenté en figure 1, la chaîne de traitements aboutissant à l'estimation de la distribution en taille des particules de dioxyde de titane ($TiO_2$) [1] comporte plusieurs étapes : la détection et la segmentation de chaque particule présente dans l'image, la classification de l'état d'agrégation de chaque particule et enfin le calcul du diamètre équivalent (diamètre de Feret ou diamètre de surface équivalent, ...). On peut également procéder à la complétion des particules partiellement visibles comme proposé par Coquelin et *al.* (2019). Bien que chaque étape doit faire l'objet d' une attention particulière, l'étape clé reste évidemment la segmentation des particules, en effet, les autres étapes dépendent directement des performances initiales de segmentation. Les sections 2 et 3 présentent les développements spécifiques réalisés pour ce cas d'étude respectivement pour les tâches de segmentation et de classification pour finir en section 4 par la présentation des performances obtenues sur les différents jeux de données de test.

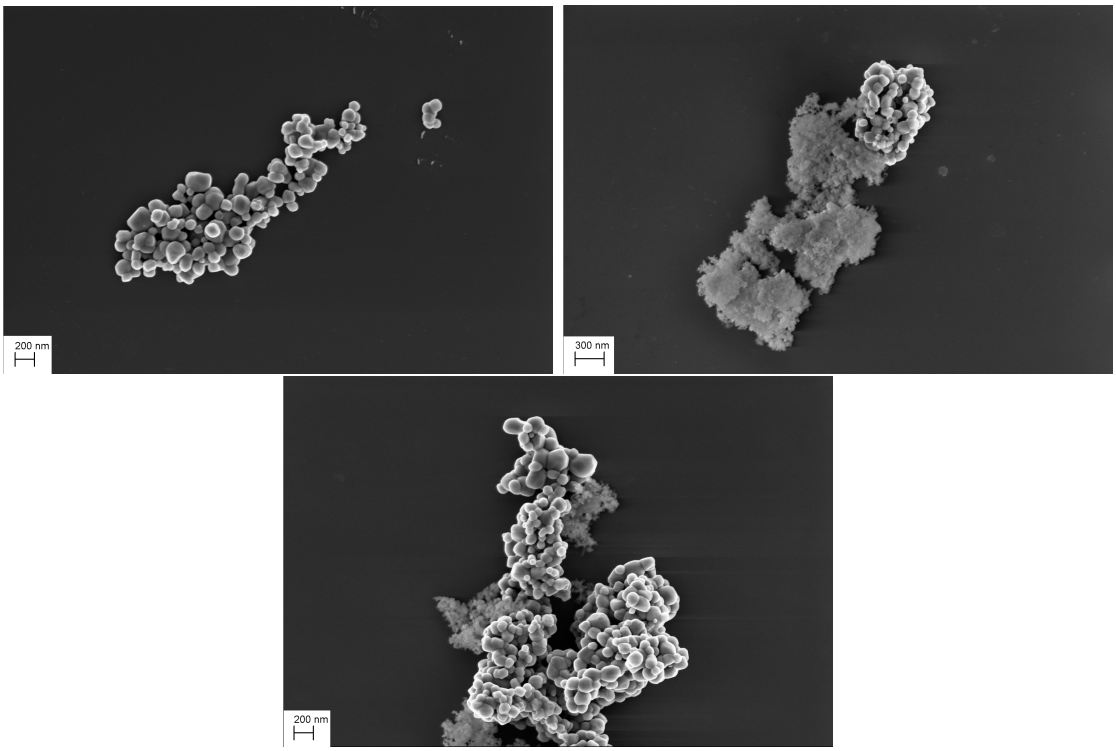



Figure 1: Exemples d'images de particules de $TiO_2$ obtenues au MEB présentant des agrégats typiques avec pour deux images la présence à l'arrière plan de particules de $SiO_2$.

[1] la distribution en taille est obtenue en estimant le diamètre équivalent de chaque particule de l'image

## 2 Segmentation des particules de $TiO_2$

**La base de données métier** La base de données est constituée de 77 images en niveaux de gris de taille 2048×1536 segmentées manuellement par des experts en nanométrologie pour un total de 5947 particules de $TiO_2$. La segmentation d'une particule par un annotateur correspond à un masque binaire.

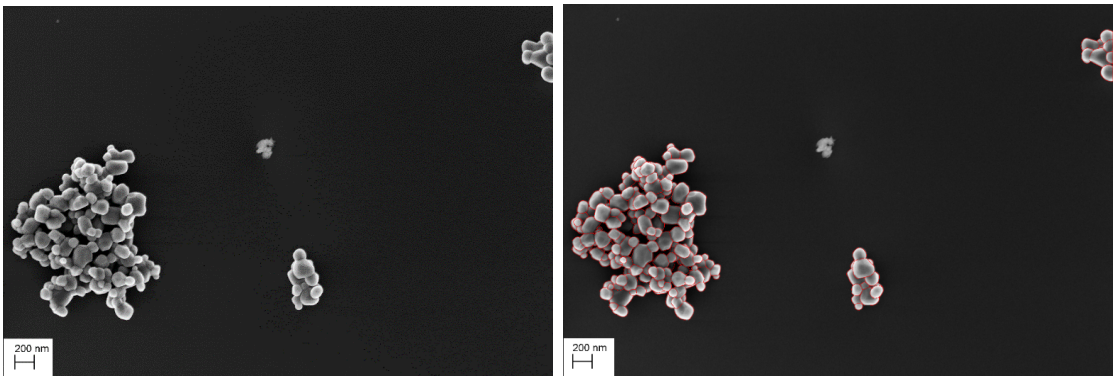



Figure 2: Exemple d'image de particules de $TiO_2$ mesurées au MEB et sa version segmentée.

**Le réseau de neurones sélectionné** L'algorithme Mask-RCNN développé par He *et al.* (2017) utilise trois réseaux de neurones : le premier (backbone network) est utilisé pour extraire des caractéristiques de l'image en entrée; le deuxième (region proposal network ou RPN) est utilisé pour générer des régions candidates contenant des objets d'intérêt[2]; le dernier réseau (mask head) prend en entrée les régions proposées et génère des masques binaires pour chaque objet détecté dans l'image. Nous utilisons un réseau pyramidal (FPN, Lin *et al.* (2017)) comme backbone afin de capturer des informations détaillées à des échelles fines ainsi que des informations contextuelles à des échelles plus larges dans l'image. Pour fusionner les caractéristiques extraites provenant de différentes résolutions spatiales, des connexions résiduelles sont employées (ResNet-50-FPN). Enfin, afin de s'assurer un haut niveau de résolution pour la génération de masques, des couches supplémentaires ont été ajoutées à l'architecture initialement proposée pour la génération de masques.

Comme souvent lorsque l'on souhaite utiliser les récents développements dans la communauté de l'apprentissage statistique (Machine Learning ou Deep Learning) dans l'industrie, la question du nombre de données disponibles peut s'avérer être un verrou à l'utilisation. Conscient du nombre limité de données à disposition, nous avons recours aux techniques dites de l'augmentation de données et du transfert d'apprentissage.

[2]il produit des boites englobantes ou "ancres" avec des scores associés en parcourant les cartes de caractéristiques extraites par le backbone

**Augmentation de données** À partir des images annotées dites de référence, chaque agglomérat/agrégat est extrait pour constituer une bibliothèque d'agglomérats/agrégats. La figure 2 illustre l'extraction de trois agrégats. L'augmentation de données consiste alors à simuler de nouvelles images en appliquant tout d'abord de manière aléatoire un retournement et une rotation à ces agrégats, puis en positionnant de manière aléatoire ces agrégats de particules sur différents fonds vides (avec ou sans matrice, la matrice pouvant être constituée d'autres particules telles que des particules de dioxyde de silice). Partant de ces images simulées, des transformations sont appliquées de façon aléatoire aux images pendant l'entraînement (flou gaussien, normalisation du contraste, bruit gaussien additif, multiplication de la valeur des pixels, ...).

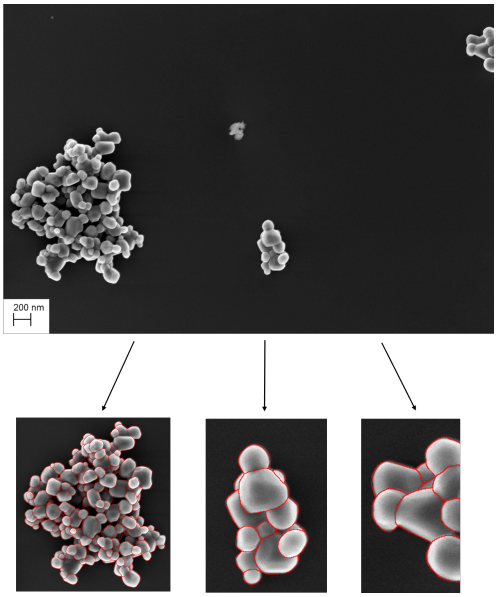



Figure 3: Image originale de particules de $TiO_2$ mesurée au MEB (figure du haut) et les agrégats annotés (3 figures du bas) qui seront utilisés pour simuler de nouvelles images.

**Transfert d'apprentissage** Le transfert d'apprentissage (Torrey et Shavlik (2010)) sert à palier au manque de données et consiste à initialiser le réseau avec les poids d'un réseau avec la même architecture mais entraîné pour une tâche différente. L'approche la plus couramment utilisée consiste à entraîner d'abord les couches finales du réseau [3], puis, après un nombre choisi d'itérations, à entraîner l'ensemble du réseau [4] pour la tâche spécifique, dans notre cas la segmentation des particules de $TiO_2$. Les poids transférés proviennent de l'algorithme entraîné sur la base de données MS COCO proposée par Lin *et al.* (2014) qui contient 91 catégories d'objets distinctes et près de 2 500 000 instances annotées sur 328 000 images.

**Ajustement des hyper-paramètres** Les images MEB étudiées comportent un grand nombre de particules de $TiO_2$ dans chaque image, parfois plusieurs plusieurs centaines, voire des milliers. Or, l'algorithme a été initialement développé pour détecter des personnes, des voitures, etc... En d'autres termes, le nombre d'instances de classes par image est drastiquement plus élevé. L'une des premières adaptations de l'algorithme a donc été réalisée pour le réseau assurant la proposition de régions (RPN). L'objectif du RPN est de parcourir toutes les cartes de caractéristiques extraites par le backbone (approche par fenêtre glissante via des éléments géométriques appelés "ancres"), d'identifier les objets et d'estimer précisément les boites englobantes pour chaque objet. Les modifications suivantes ont été apportées afin de prendre en compte la spécificité de la tâche de segmentation : les tailles des ancres ont été modifiées en fonction de la taille minimale et maximale des particules de $TiO_2$ (8, 16, 32,

[3] les poids transférés dans les premières couches restent alors fixes
[4] l'ensemble des poids du réseau sont mis à jour

64 et 96 pixels); le nombre d'ancres entraînées a été adapté par rapport au nombre maximal de particules dans une image (1024); l'écart entre deux ancres consécutives a été fixé à (1) en raison du phénomène d'agrégation des particules.

**Entraînement du réseau de neurones** Le jeu de données d'entraînement comporte 699 images, dont 77 images réelles et 622 images simulées via la stratégie d'augmentation de données présentée précédemment. La stratégie d'entraînement reste très artisanale à ce stade et a été purement guidée par les performances constatées. Les têtes du réseau ont été entraînées pendant 38 itérations[5] avec un taux d'apprentissage de 0.001 puis 4 itérations avec un taux d'apprentissage de 0.0001. L'ensemble du réseau a ensuite été entraîné pendant 28 itérations avec un taux d'apprentissage de 0.001 et enfin 7 itérations avec un taux d'apprentissage de 0.0001. L'algorithme d'optimisation choisi est un algorithme de descente de gradient stochastique (SGD) avec avec un momentum (0.9) et un seuillage de la norme du gradient (5.0). Comme suggéré par Krogh et Hertz (1992), nous utilisons la régularisation $L_2$ avec une décroissance de poids (0.0001). Le réseau a été entraîné sur une simple carte graphique GPU NVIDIA GeForce RTX 2080 avec 8 Go de mémoire.

# 3 Classification de l'état d'agrégation des particules de $TiO_2$

Cette étape de classification fait office de sélection des données utiles pour l'estimation de la distribution en taille pour laquelle seules les particules entièrement imagées doivent être mesurées. Les particules détectées sont donc réparties au sein des 5 classes suivantes : Isolée (*la particule est complètement imagée et située en dehors d'un agrégat*), Complète (*la particule est complètement imagée et située dans un agrégat*), Complète* (*la particule est complètement imagée mais imbriquée dans une autre particule, c'est l'état intermédiaire entre "complète" et masquée"*), Masquée (*la particule est en partie cachée par une autre particule*), Inutilisable (*la particule est masquée par une autre particule dont la surface visible est très réduite*). Le tableau 1 illustre les différentes classes en détaillant le nombre de particules par classe pour les jeux de données d'entraînement et de test.

**La base de données** La base de données a été complètement crée à partir des prédictions de l'algorithme de segmentation sur les images de référence. Des experts en nanométrologie ont ensuite statué sur l'état d'agrégation de chaque particule à partir des vignettes extraites. Pour rappel, une segmentation correspond à un masque binaire. Les images d'entrée sont ici composées de deux canaux : le premier contient l'image de la particule en nuance de gris et le deuxième contient le masque de segmentation produit lors de l'étape précédente. Nous avons ajouté la classe "A jeter" pour les particules mal segmentées.

[5]le terme itération ici fait référence à un seul passage complet à travers l'ensemble des données d'entraînement lors de l'apprentissage du modèle (epoch en anglais)

| Vignette | Etat d'agrégation | # (Entraînement) | # (Test) |
|---|---|---|---|
|  | Isolée | 567 | 54 |
|  | Complète | 1129 | 98 |
|  | Complète* | 2329 | 250 |
|  | Masquée | 4611 | 478 |
|  | Inutilisable | 1747 | 183 |
| | À jeter | 2789 | 320 |
| | **Total** | 13172 | 1383 |

Table 1: Les différents états d'agrégation des particules de $TiO_2$ et le nombre de particules présentes dans les jeux de données d'entraînement et de test pour la tâche de classification.

**Le réseau de neurones sélectionné** La figure 4 présente le schéma du réseau VGG16 utilisé, initialement développé par Simonyan et Zisserman (2014). La philosophie de cette architecture consiste en l'enchaînement de blocs convolutifs munis de fonction d'activation ReLU, chacun séparé par une couche de max pooling permettant de réduire la résolution spatiale. Dans une architecture VGG, lorsque la résolution spatiale est réduite par 2 (en hauteur et en largeur, soit 4 fois moins de pixels), on augmente par 2 la profondeur de l'image (2 fois plus de filtres).

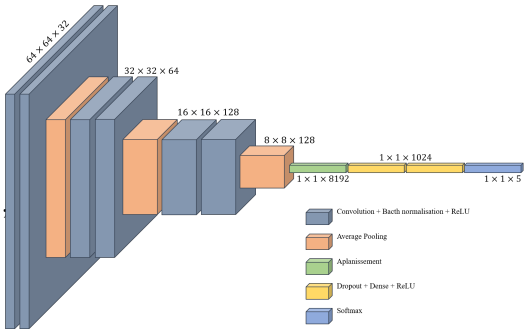



Figure 4: Architecture du réseau VGG16 utilisé pour la tâche de classification de l'état d'agrégation des particules de $TiO_2$.

Les dernières couches complètement connectées, suivi d'une couche softmax (régression logistique multinomiale) servent ensuite de classifieur. L'entraînement de ce type de réseau est extrêmement classique et ne sera pas détaillé ici.

# 4 Résultats

Dans cette section, nous présentons les résultats obtenus respectivement par le Mask-RCNN modifié (section 2) et par le modèle VGG16 (section 3) pour les tâches de segmentation et de classification de l'état d'agrégation des particules de $TiO_2$. Il est à préciser qu'il n'existe pas aujourd'hui dans la littérature d'approches concurrentes complètement automatisées auxquelles se comparer. Pour cette raison, les comparaisons numériques proposées

considèrent l'opérateur humain comme la référence.

## 4.1 Segmentation

Le jeu de données de test est composé de 19 images pour un nombre total de 3741 particules de $TiO_2$. Ces 19 images couvrent plusieurs configurations rencontrées régulièrement en pratique : particules de $TiO_2$ agrégées, dispersées ou mélangées avec d'autres types de particules.

**performance en détection** Il s'agit simplement ici d'évaluer sur le jeu de test le nombre de particules détectées par l'algorithme de segmentation présenté en section 2 en fonction du statut d'agrégation des particules. Le tableau 2 montre que la détection des particules d'intérêt pour l'estimation de la distribution en taille est satisfaisante ($> 99\%$ pour les particules de la classe Complète et $> 96\%$ pour les particules de la classe Complète*).

| | **Total** | **Complète** | **Complète*** | **Masquée** | **Inutilisable** |
|---|---|---|---|---|---|
| Référence | 3741 | 341 | 515 | 1302 | 1583 |
| Mesure | 3135 | 339 | 495 | 1253 | 1048 |
| % bonnes détections | 83.80 | 99.41 | 96.11 | 96.23 | 66.2 |

Table 2: Nombre de particules détectées sur le jeu de test en fonction de l'état d'agrégation des particules.

**performance en segmentation** Évaluer les masques prédits pour chaque particule nécessite une mesure de similarité entre le masque de référence et le masque généré par l'algorithme. Il existe un grand nombre de mesures dans la littérature (indice de Jaccard (1901), indice de Tversky (1977), ...), nous faisons le choix du coefficient de Sørensen–Dice ou coefficient de similarité de Dice (développé indépendamment par Dice (1945) et Sorensen (1948)), il est donné par la formule suivante :

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|}$$

, où A et B désignent respectivement les masques binaires de référence et prédit et $|A|$ correspond la taille de A, autrement dit le nombre de pixels ayant pour valeur 1 dans le masque A. Une parfaite concordance entre les deux masques se traduit par un coefficient de 1, tandis qu'une valeur de 0 indique une absence de chevauchement entre les masques. Le tableau 3 présente les paramètres statistiques de la distribution du coefficient de Sørensen–Dice sur les particules détectées sur le jeu de test en fonction de l'état d'agrégation. L'analyse des résultats est similaire à celle pour la détection, c'est-à-dire que l'algorithme prédit des masques similaires aux masques de référence pour les particules d'intérêt ($DSC$ moyen $> 0.95$ pour les particules des classes Complète et Complète*).

| | *DSC* | | | | |
|---|---|---|---|---|---|
| | **Moyenne** | **Médiane** | **Ecart-type** | **Min** | **Max** |
| Total | 0.936 | 0.948 | 0.040 | 0.750 | 0.989 |
| Complète | 0.953 | 0.959 | 0.024 | 0.819 | 0.987 |
| Complète* | 0.953 | 0.958 | 0.022 | 0.858 | 0.986 |
| Masquée | 0.949 | 0.957 | 0.026 | 0.765 | 0.989 |
| Inutilsable | 0.903 | 0.913 | 0.046 | 0.756 | 0.975 |

Table 3: Statistiques du coefficient de Sørensen–Dice sur les particules détectées sur le jeu de test.

**estimation du diamètre de surface équivalent**
Les mesures de performance se sont intéressées jusqu'ici à la détection des particules et à la prédiction des masques de segmentation. Pour obtenir la distribution en taille d'un échantillon, l'étape finale consiste à calculer le diamètre équivalent de chaque particule. Ici nous nous intéressons au diamètre de surface équivalent. La figure 5 montre la distribution des résidus en pourcentage entre le diamètre de surface équivalent calculé à partir de la segmentation de référence et la segmentation prédite automatiquement. Les résultats révèlent que 96 % des mesures sur le jeu de test présentent moins de 5 % d'erreur sur l'estimation du diamètre et également que 51 % des mesures présentent moins de 1 % d'erreur.

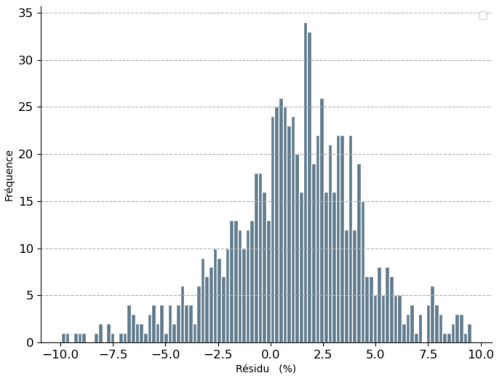



Figure 5: Histogramme des résidus (en pourcentage) pour le diamètre de surface équivalent sur le jeu de test.

## 4.2 Classification

Sur le jeu de test détaillé en section 3, le VGG16 utilisé atteint un score de bonne classification de 76.3% (soit 23.8% de mauvaises classifications). Bien que ce score puisse paraître assez faible, il est du même ordre de grandeur que la performance estimée sur un opérateur humain, ce qui en fait par conséquent un bon classifieur. En effet, il peut être très complexe de différentier les particules des classes masquées/inutilisables ainsi que les particules des classes complète/complète*/masquée comme indiqué par le tableau 4 présentant la matrice de confusion des prédictions.

Le tableau 5 récapitule le nombre de bonnes et mauvaises classifications en fonction du seuil choisi sur la prédiction du modèle. Bien évidemment, augmenter ce seuil conduit à réduire considérablement le nombre de particules (97.2% de bonnes détections pour un seuil de 0.95 qui conduit à conserver 394 particules pour un nombre initial de 1383, soit 28.5%

| **Vraie classe** | **Classe prédite** | | | | | |
|---|---|---|---|---|---|---|
| | Complète | A jeter | Isolée | Masquée | Complète* | Inutilisable |
| Complète | 77 | 1 | 0 | 1 | 19 | 0 |
| A jeter | 2 | 251 | 1 | 34 | 9 | 23 |
| Isolée | 0 | 2 | 52 | 0 | 0 | 0 |
| Masquée | 4 | 15 | 0 | 376 | 59 | 24 |
| Complete* | 30 | 8 | 0 | 32 | 177 | 3 |
| Inutilisable | 0 | 14 | 0 | 42 | 5 | 122 |

Table 4: Matrice de confusion de la tâche de classification par VGG16.

seulement des particules conservées). L'analyse de ces résultats conduit à sélectionner un seuil de 0.80 qui conduit à 88.7% de bonnes classifications en maintenant plus de 50 % des particules.

| | **Seuil** | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.0 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.95 |
| Complète | 77 (78.6) | 71 (78.0) | 62 (84.9) | 56 (84.5) | 50 (89.3) | 41 (95.3) | 33 (97.1) |
| A jeter | 251 (78.4) | 242 (81.8) | 234 (83.9) | 221 (86.0) | 203 (91.0) | 181 (95.8) | 160 (98.8) |
| Isolée | 52 (96.3) | 52 (96.3) | 52 (96.3) | 52 (96.3) | 52 (96.3) | 52 (96.3) | 52 (96.3) |
| Masquée | 376 (78.7) | 359 (75.1) | 324 (83.7) | 273 (85.0) | 224 (89.6) | 163 (95.9) | 97 (0.97) |
| Complete* | 177 (70.8) | 165 (71.7) | 147 (76.2) | 118 (80.3) | 86 (80.4) | 37 (80.4) | 19 (86.4) |
| Inutilisable | 122 (66.7) | 117 (70.5) | 98 (76.0) | 85 (77.3) | 71 (84.5) | 44 (91.7) | 22 (95.7) |
| **Total** | 1055 (76.3) | 1006 (78.3) | 917 (82.3) | 805 (84.6) | 686 (88.7) | 518 (94.4) | 383 (97.2) |
| **# particules** | 1383 (100.0) | 1284 (92.8) | 1114 (80.5) | 952 (68.8) | 773 (55.9) | 549 (39.7) | 394 (28.5) |

Table 5: Nombre de bonnes classifications en fonction de la classe de la particule et du score de prédiction (les pourcentages sont en ()). # particules précise le nombre de particules restantes après seuillage.

# 5 Conclusion

L'automatisation de la caractérisation de la taille des particules de dioxyde de titane ($TiO_2$) obtenues par MEB devient souhaitable compte-tenu du temps et des ressources nécessaires à l'heure actuelle pour réaliser cette tâche, souvent réalisée manuellement, qui est rendue difficile par la non sphéricité des particules et à leur tendance à fortement s'agréger. Recourir à l'apprentissage statistique malgré une base de données métier de taille réduite est possible aujourd'hui en utilisant les techniques de transfert d'apprentissage et d'augmentation de données. Les résultats obtenus sont convaincants avec notamment 96 % des mesures sur le jeu de test présentant moins de 5 % d'erreur sur l'estimation du diamètre de surface équivalent. Ces résultats ouvrent clairement la voie à l'automatisation complète de ce type de caractérisation purement basée sur la vision par ordinateur, tâche sur laquelle ces algorithmes

ont montré leur supériorité depuis quelques années. La suite des travaux portera sur la quantification de l'incertitude associée aux prédictions des réseaux de neurones utilisés dans cette chaîne de traitements automatisés.

# Bibliographie


Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), pp. 297-302.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske skrifter*, 5, pp. 1-34.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société vaudoise des sciences naturelles*, 37, pp. 241-272.

Tversky, Amos (1977). Features of Similarity. *Psychological Review*. 84 (4), pp. 327–352.

Weir, A., Westerhoff, P., Fabricius, L., Hristovski, K., et Von Goetz, N. (2012). Titanium dioxide nanoparticles in food and personal care products. *Environmental science and technology*, 46(4), pp. 2242-2250.

Hwang, J. S., Yu, J., Kim, H. M., Oh, J. M., et Choi, S. J. (2019). Food additive titanium dioxide and its fate in commercial foods. *Nanomaterials*, 9(8), 1175.

Torrey, L., et Shavlik, J. (2010). Transfer learning. *In Handbook of research on Machine Learning applications and trends: algorithms, methods, and techniques*, pp. 242-264. IGI global.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... et Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European Conference in Computer Vision–ECCV 2014*, 13, pp. 740-755.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., et Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125.

He, K., Gkioxari, G., Dollár, P., et Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969.

Krogh, A., et Hertz, J. (1991). A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, pp. 950-957.

Simonyan, K., et Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

# Sous-échantillonnage de données pour les réseaux de neurones bayésiens

Eiji Kawasaki [1] & Markus Holzmann [2] & Lawrence Adu-Gyamfi [1]

[1] *Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France*
[2] *Univ. Grenoble Alpes, CNRS, LPMMC, 38000 Grenoble, France*

**Résumé.** La mise au point d'une méthode efficace de quantification d'incertitude en apprentissage profond est un tâche importante mais difficile car elle implique le calcul de la distribution prédictive en marginalisant l'ensemble des paramètres des réseaux de neurones. Dans ce contexte, les algorithmes Monte Carlo par chaînes de Markov ne s'adaptent pas bien aux grands volumes de données, ce qui entraîne des difficultés dans l'échantillonnage des distributions a posteriori des réseaux de neurones. En visant cet objectif d'inférence bayésienne, nous proposons de montrer qu'une généralisation de l'algorithme Metropolis-Hastings permet de restreindre l'évaluation de la vraisemblance à des sous-ensembles des données d'entraînement nommés mini-lot. Comme cette méthode nécessite le calcul d'une "pénalité de bruit" déterminée par la variance de la fonction de perte sur ces mini-lots, nous appelons cette stratégie de sous-échantillonnage des données "réseaux neuronaux bayésiens à pénalité de bruit".

**Mots-clés.** Réseaux de neurones bayésiens, Monte Carlo par chaînes de Markov

**Abstract.** The development of an effective uncertainty quantification method that computes the predictive distribution by marginalizing over Deep Neural Network parameter sets remains an important, challenging task. In this context, Markov Chain Monte Carlo algorithms do not scale well for large datasets leading to difficulties in Neural Network posterior sampling. We show that a generalization of the Metropolis Hastings algorithm allows to restrict the evaluation of the likelihood to small mini-batches in a Bayesian inference context. Since it requires the computation of a so-called "noise penalty" determined by the variance of the training loss function over the mini-batches, we refer to this data subsampling strategy as Penalty Bayesian Neural Networks.

**Keywords.** Bayesian Neural Network, Markov Chain Monte Carlo

# 1 Réseaux de neurones bayésien à pénalité de bruit

## 1.1 Introduction

Le développement d'une méthode efficace de quantification de l'incertitude de prédictions de modèles de réseaux de neurones profonds est une tâche importante et difficile [1]. Les méthodes d'inférence bayésienne permettent d'obtenir la distribution a posteriori des paramètres

en utilisant par exemple l'inférence variationnelle ou bien encore l'échantillonnage Monte Carlo. Les techniques de Monte Carlo par chaînes de Markov (MCMC) sont généralement considérées comme la référence en matière d'inférence bayésienne [2]. Cependant, l'exploration de l'espace des paramètres d'un réseau de neurones par une chaîne de Markov ne s'adapte pas bien aux grands volumes de données. En effet, elle nécessite l'évaluation de la log-vraisemblance du modèle sur l'ensemble des données à chaque étape d'itération. En pratique, cela constitue un grave obstacle à l'utilisation de l'échantillonnage des réseaux de neurones bayésiens. Le développement d'un algorithme d'échantillonnage MCMC pour des réseaux de neurones bayésiens capables de traiter des ensembles de données aux dimensions satisfaisantes dans le cadre de l'apprentissage profond reste donc un problème ouvert.

Par analogie avec les techniques de descente de gradient stochastique omniprésentes en apprentissage automatique, nous proposons une stratégie de sous-échantillonnage des données pour l'évaluation de la distribution a posteriori d'un réseau de neurones. Cela nous conduit à une variante du MCMC que nous appelons réseau de neurones bayésien avec pénalité de bruit. En effet, ne pas prendre en compte le bruit dû au sous-échantillonnage des données ne permet pas de correctement approcher la distribution par MCMC. Ce constat a été établi dans le contexte de l'inférence bayésienne : plusieurs méthodologies de sous-échantillonnage MCMC ont été proposées pour généraliser l'algorithme de Metropolis-Hastings et maîtriser ce biais [3]. Nous montrons, à la fois théoriquement et empiriquement, que notre approche originale permet un échantillonnage non baisé de la loi a posteriori en calculant explicitement la variance de la différence des fonctions de perte d'un mini-lot de données.

## 1.2 Distribution a posteriori et mini-lots

Nous considérons ici un vecteur $\theta$ qui décrit les paramètres d'un modèle, ce vecteur pourra notamment représenter les poids et les biais d'un réseau de neurones. Nous définissons $p(\theta)$ comme une distribution a priori sur cet ensemble de paramètres. Les distributions a priori usuelles en apprentissage profond sont les distributions gaussienne et de Laplace, qui correspondent respectivement aux régularisations L2 et L1. Dans le contexte de l'apprentissage supervisé, nous appelons $p(y|x,\theta)$ la probabilité d'une cible $y$ compte tenu d'une donnée d'entrée $x$ et d'un vecteur de paramètres $\theta$. L'incertitude sur les paramètres $\theta$ étant donné un ensemble d'observations $\mathcal{D}$ est décrite par la distribution a posteriori qui est définie suivant,

$$p(\theta|\mathcal{D}) = \frac{p(\theta)\prod_{i=1}^{N} p(y_i|x_i,\theta)}{p(\mathcal{D})} \tag{1}$$

où $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^{N}$. A une constante près, $p(\theta|\mathcal{D}) \propto e^{-\mathcal{L}_{\mathcal{D}}(\theta)}$ où la fonction de perte $\mathcal{L}_{\mathcal{D}}(\theta)$ correspond au log négatif de la distribution a posteriori

$$\mathcal{L}_{\mathcal{D}}(\theta) = -\log p(\theta) - \sum_{i=1}^{N} \log p(y_i|x_i,\theta) \tag{2}$$

Le dernier terme est la négative log-vraisemblance. Ce choix de fonction de perte n'est donné qu'à titre d'illustration et ne réduit pas la généralité de l'approche présentée ici, car nous

aurions également pu envisager une configuration non supervisée dans laquelle $\mathcal{D} = \{(x_i)\}_{i=1}^N$ et $\mathcal{L}_\mathcal{D}(\theta) = -\log p(\theta) - \sum_{i=1}^N \log p(x_i|\theta)$.

En apprentissage conventionnel, les paramètres du réseau sont généralement estimés par une descente de gradient stochastique qui cible le maximum de la distribution a posteriori en utilisant des mini-lots de données MB. Il en résulte une fonction de perte définie comme suit

$$\mathcal{L}_{\text{MB}}(\theta) = -\log p(\theta) - \frac{N}{n}\sum_{i=1}^{n} \log p(y_i|x_i, \theta) \tag{3}$$

où $n$ correspond à la taille du mini-lot qui contient des données sous-échantillonnées sans remise, de sorte que par définition $\langle \mathcal{L}_{\text{MB}}(\theta) \rangle = \mathcal{L}_\mathcal{D}(\theta)$.

## 1.3 Metropolis-Hastings et bruit gaussien

Il est possible d'obtenir un ensemble i.i.d. d'échantillons de la distribution a posteriori définie dans l'équation 1 à l'aide d'un algorithme MCMC en explorant l'espace de définition de $\theta$ à l'aide de chaînes de Markov. Il est bien connu que l'équation de l'équilibre détaillé est une condition suffisante mais non nécessaire garantissant que ce processus de Markov possède une distribution stationnaire correspondant à l'équation 1. L'équilibre détaillé est donné par

$$A(\theta, \theta')q(\theta|\theta')e^{-\Delta(\theta',\theta)} = A(\theta', \theta)q(\theta'|\theta) \tag{4}$$

où $A(\theta', \theta)$ correspond à la probabilité d'accepter un déplacement d'un vecteur de paramètres $\theta$ vers $\theta'$. Ce changement d'état est suggéré par la distribution $q(\theta'|\theta)$. Par souci de concision, nous considérons dans la suite une distribution de proposition symétrique $q(\theta'|\theta) = q(\theta|\theta')$. En utilisant l'algorithme de Metropolis-Hastings, l'acceptation s'écrit alors $A(\theta', \theta) = \min\left(1, e^{-\Delta(\theta',\theta)}\right)$ où $\Delta(\theta', \theta)$ correspond à la différence de fonctions de perte définie par

$$\Delta(\theta', \theta) = \mathcal{L}_\mathcal{D}(\theta') - \mathcal{L}_\mathcal{D}(\theta) \tag{5}$$

Nous souhaitons calculer les différences de fonction de perte sur des mini-lots aléatoires plutôt que sur l'ensemble des données $\mathcal{D}$. Par conséquent, nous introduisons une variable aléatoire $\delta(\theta', \theta)$ qui fournit une estimation non biaisée de $\Delta(\theta', \theta)$

$$\delta(\theta', \theta) \sim \mathcal{N}(\Delta(\theta', \theta), \sigma^2(\theta', \theta)) \tag{6}$$

ce qui signifie que nous pouvons écrire $\delta(\theta', \theta)$ comme étant égal à la différence de perte $\Delta(\theta', \theta)$ à laquelle on ajoute un bruit dont nous faisons l'hypothèse qu'il est distribué selon une gaussienne. Généralement, l'augmentation de la taille $n$ des mini-lots diminue l'amplitude du bruit, c'est-à-dire qu'il réduit la variance $\sigma^2(\theta', \theta)$.

Comme le montre la figure 1, l'estimateur de la différence des fonctions de perte $\delta(\theta', \theta)$ empêche les algorithmes MCMC d'échantillonner la distribution a posteriori ciblée si le bruit qu'il introduit n'est pas correctement pris en compte. Dans le contexte de la physique statistique et de la chimie informatique, Ceperley et Dewing (1999) [5] ont généralisé l'algorithme de marche aléatoire de Metropolis-Hastings à la situation où la différence $\Delta$ (différence

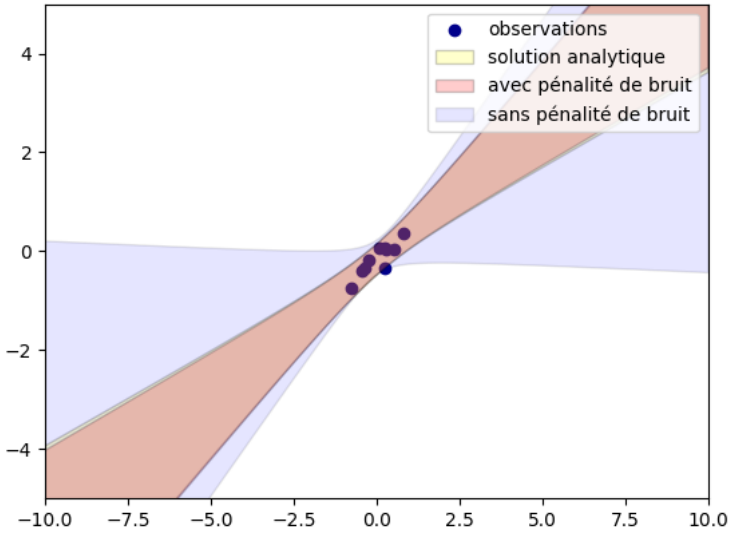



Figure 1: Tracé des distributions prédictives a posteriori calculées pour une régression linéaire univariée. Les zones colorées correspondent à la moyenne des distributions ± un écart-type. La courbe bleue est calculée en remplaçant naïvement $\Delta$ par $\delta$ dans l'algorithme MH. La différence bruitée $\delta$ est calculée sur un seul mini-lot contenant un sous-ensemble de 2 données. La courbe jaune inclut le terme de pénalité de bruit supplémentaire tel que défini dans l'eq 7. Nous remarquons qu'elle se superpose à la courbe rouge qui représente la solution analytique d'une régression linéaire bayésienne avec une distribution a priori gaussienne et une variance aléatorique connue [4].

d'énergies dans leur cas d'étude) est bruitée par un bruit gaussien et ne peut être qu'estimée. Ils ont montré qu'il est possible d'échantillonner la distribution ciblée malgré la présence d'un bruit. Pour cela, il est nécessaire de modifier la probabilité d'acceptation et d'appliquer une pénalité de bruit $-\sigma^2(\theta', \theta)/2$ à la différence de perte dans le ratio d'acceptation $A$, de telle sorte que :

$$A(\delta, \theta', \theta) = \min\left(1, e^{-\delta(\theta',\theta)-\sigma^2(\theta',\theta)/2}\right) \tag{7}$$

On peut alors montrer que l'équilibre détaillé est satisfait en moyenne suivant l'équation 8, ce qui est une condition suffisante pour que la chaîne de Markov échantillonne la distribution sans biais dans le régime stationnaire.

$$\begin{aligned} &\int d\delta A(\delta, \theta, \theta') q(\theta|\theta') \mathcal{N}(\delta; \Delta(\theta', \theta), \sigma^2(\theta', \theta)) e^{-\delta} \\ &= \int d\delta A(\delta, \theta', \theta) q(\theta'|\theta) \mathcal{N}(\delta; \Delta(\theta, \theta'), \sigma^2(\theta, \theta')) \end{aligned} \tag{8}$$

Notons que cette méthode de pénalité de bruit peut être étendue à une distribution de proposition non symétrique $q(\theta'|\theta)$. A ce titre, les réseaux de neurones bayésiens sont

souvent échantillonnés par dynamique de Langevin [6] ou encore par Monte Carlo hybride [7]. L'inconvénient de la pénalité de bruit est qu'elle entraîne une diminution exponentielle de l'acceptation $A(\delta, \theta, \theta')$, puisque la variance $\sigma^2(\theta', \theta)$ est toujours non négative. Notons en outre que dans le cas de l'échantillonnage a posteriori, $\sigma^2(\theta', \theta)$ n'est en général pas connu et ne peut qu'être estimé. Il est possible d'étendre ce raisonnement pour prendre en compte des variances bruitées [5].

## 1.4 La pénalité de bruit en pratique

Afin d'obtenir une acceptation moyenne raisonnable $A(\delta, \theta', \theta_t)$, c'est-à-dire qui ne tend pas vers 0, la différence de fonction de perte $\delta(\theta', \theta_t)$ doit dominer la variance $\sigma^2(\theta', \theta_t)/2$ qui est par définition toujours positive. Toutefois, en pratique, la pénalité de bruit domine souvent tout gain entre $\theta_t$ et $\theta'$ si cette différence n'est calculée que sur un seul petit mini-lot. Cela conduit à une dimininution exponentielle de l'acceptation et à de longs temps de corrélation de la chaîne de Markov.

Pour éviter cette situation, nous définissons $\delta(\theta', \theta)$ comme une moyenne empirique de la différence des fonctions de perte :

$$\delta(\theta', \theta) = \frac{1}{M} \sum_{j=1}^{M} \left( \mathcal{L}_{\mathrm{MB},j}(\theta') - \mathcal{L}_{\mathrm{MB},j}(\theta) \right) \tag{9}$$

où $\mathrm{MB}, j$ correspond à un mini-lot $j$ choisi aléatoirement. Par définition, la moyenne est un estimateur non biaisé tel que $\langle \delta(\theta', \theta) \rangle = \Delta(\theta', \theta)$. Nous remarquons ici que le théorème central limite garantit que $\delta(\theta', \theta)$ est distribué selon une gaussienne dans la limite d'un nombre $M$ infini.

La variance $\sigma^2$ de la variable aléatoire $\delta(\theta', \theta)$ diminue lorsque le nombre de mini-lots $M$ augmente. Cette variance théorique nous est inconnue, mais nous pouvons en calculer une estimation non biaisée :

$$\sigma^2(\theta', \theta) \simeq \frac{1}{M(M-1)} \sum_{j=1}^{M} \left( \mathcal{L}_{\mathrm{MB},j}(\theta') - \mathcal{L}_{\mathrm{MB},j}(\theta) - \delta(\theta', \theta) \right)^2 \tag{10}$$

La figure 2 illustre une expérience numérique d'un réseaux de neurones bayésien à pénalité de bruit sur une tâche de régression à partir de données synthétiques. L'erreur sur l'estimation de la variance $\sigma^2(\theta', \theta)$ n'est pas prise en compte, nous prenons ici l'hypothèse que ses variations en fonction de $\theta'$ et $\theta$ dominent largement le bruit dû à son estimation. Les corrections d'ordre supérieur tenant compte de cette incertitude sont discutées dans [5].

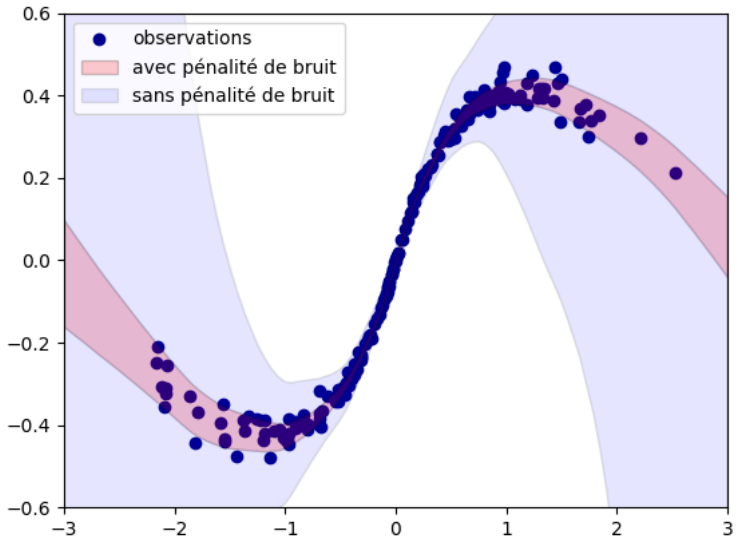



Figure 2: Tracé des distributions prédictives a posteriori calculées pour un problème de régression synthétique univariée. Le modèle de vraissemblance est une gaussienne dont la moyenne et la variance sont paramétrisés par un réseau de neurones [8].

# Bibliographie

[1] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A Survey of Uncertainty in Deep Neural Networks. *arXiv:2107.03342 [cs, stat]*, July 2021. arXiv: 2107.03342.

[2] Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? *arXiv:2104.14421 [cs, stat]*, April 2021. arXiv: 2104.14421.

[3] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

[4] Christopher M Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[5] D. M. Ceperley and M. Dewing. The penalty method for random walks with uncertain energies. *The Journal of Chemical Physics*, 110(20):9812–9820, May 1999.

[6] Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. page 8.

[7] Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1996.

[8] Christopher M. Bishop. Mixture density networks, 1994. Num Pages: 26 Place: Birmingham Publisher: Aston University.

# Données fonctionnelles 1

# ACP pour données fonctionnelles discrétisées, estimation minimax et contraintes spectrales

Nassim Bourarach [1] & Franck Picard [2]& Vincent Rivoirard [1] & Angelina Roche [1]

[1] *CEREMADE, Université Paris-Dauphine PSL , nassim.bourarach@dauphine.psl.eu, Vincent.Rivoirard@dauphine.fr, roche@ceremade.dauphine.fr*
[2] *LBMC, ENS de Lyon, franck.picard@ens-lyon.fr*

**Résumé.** Dans cette présentation, nous explorerons le problème de l'estimation des fonctions propres et valeurs propres de l'opérateur de covariance d'un échantillon de données fonctionnelles discrétisées et bruitées. L'analyse de l'impact de la discrétisation et du bruit sur l'inférence se fera grâce à une double asymptotique : en $n$, le nombre de courbes observées, et $p$, la taille de la grille de discrétisation.
Nous aborderons le problème pour une classe de processus très large (trajectoires $m$-Höldériennes avec $m \in \mathbb{R}^*$) et expliciterons le rôle de la régularité dans les bornes. Ces nouvelles bornes inférieures minimax seront accompagnées d'un 'nouvel' estimateur non-paramétrique, fondé sur des ondelettes, qui est optimal sans nécessiter de régularisation. Après avoir présenté le cadre et nos résultats, nous discuterons plus en détail des contraintes spectrales nécessaires à travers des illustrations et des résultats d'inconsistance.

**Mots-clés.** données fonctionnelles, ACP, borne minimax, estimation non-paramétrique

**Abstract.** In this presentation, we will explore the problem of estimating the eigenfunctions and eigenvalues of the covariance operator of a sample of discretised and noisy functional data. The impact of discretisation and noise on inference will be analysed through a double asymptotic: in $n$, the number of observed curves, and $p$, the size of the discretisation grid. We will address the problem for a very broad class of processes ($m$-Hölderian trajectories with $m \in \mathbb{R}_+^*$) and specify the role of regularity in the bounds. These new minimax lower bounds will be accompanied by a 'new' non-parametric estimator, based on wavelets, that is optimal without the need for regularisation. After presenting the framework and the main results, we will discuss the necessary spectral constraints in more details by means of inconsistency results and illustrations.

**Keywords.** functional data, PCA, minimax bound, non-parametric estimation

# 1 Introduction

## 1.1 Les données fonctionnelles

Pour faire face à l'afflux de données de plus en plus massives, beaucoup d'approches statistiques différentes ont vu le jour. Dans de nombreuses situations, les données dont on dispose se présentent comme des vecteurs d'observations en grande dimension qui cachent une dépendance importante entre les différentes entrées du vecteur. Il peut s'agir d'une dépendance spatiale ou temporelle par exemple. Ces dépendances peuvent d'une certaine façon être encapsulées dans une modélisation fonctionnelle des observations, à travers des contraintes de régularité. Ainsi, plutôt que de considérer qu'on observe des vecteurs aléatoires en très grande dimension, on considère qu'on observe des réalisations (potentiellement bruitées) d'une variable aléatoire $X_i \overset{i.i.d.}{\sim} X$ à valeur dans un espace de fonctions $\mathcal{H} \subset \mathbb{L}_2([0,1])$ l'espace des fonctions de carré intégrable sur $[0,1]$. C'est notamment le point de vue adopté dans de nombreux ouvrages dédiés aux données fonctionnelles (Ferraty et Vieu (2006), Ramsay et Silverman (2010), Hsing et Eubank (2015)).

Ce premier saut conceptuel nous éloigne de la réalité des données (on n'observe jamais plus que des vecteurs) mais permet l'accès à la multitude de résultats relatifs aux espaces de fonctions. Nos travaux s'inscrivent alors dans une seconde perspective où l'on essaie de réconcilier l'approche fonctionnelle avec les données observées, par la prise en compte de l'aspect discret des données. Les nouveaux résultats que nous obtenons prennent place dans le prolongement de ceux de Belhakem et al. (2022).

De ce fait, en ayant $p, n$ deux entiers naturels strictement positifs, nos données consistent en $p$ évaluations bruitées de $n$ réalisations de fonctions aléatoires sur une grille $(t_j)_{j=1}^p \in [0,1]$ :

$$Y_i(t_j) = X_i(t_j) + \varepsilon_{i,j}, \qquad (i,j) \in [\![1,n]\!] \times [\![1,p]\!]. \tag{1}$$

où les $\varepsilon_{i,j}$ sont i.i.d. gaussiennes, centrées, de variance $\sigma^2 > 0$ fixée, indépendantes des $X_i$ qui sont aussi i.i.d. .

## 1.2 Le problème

L'Analyse en Composantes Principales fonctionelles (ACPf), tout comme son équivalent vectoriel, joue un rôle important, à la fois comme outil de réduction de la dimension ou d'analyse exploratoire des données fonctionnelles. En pratique c'est ce qui est observé dans les papiers de Viviani et al. (2005) sur des données fMRI ou plus récemment par Warmenhoven et al. (2021) pour des données issues de la biomécanique.

Soit $D$ une dimension fixée, l'ACPf consiste à trouver un système de $D$ fonctions $(\psi_1^*, \ldots, \psi_D^*)$ minimisant l'écart quadratique entre une donnée fonctionnelle $X$ et sa projection orthogonale $\Pi_{S_D} X$ sur l'espace $S_D^* = \text{Vect}\{\psi_1^*, \ldots, \psi_D^*\}$ autrement dit

$$S_D^* \in \arg \min_{\substack{S \text{ s.e.v. de } \mathbb{L}_2([0,1],\\ \dim(S)=D}} E\left[\|X - \Pi_S X\|^2\right], \tag{2}$$

où $\Pi_S X$ est la projection orthogonale sur l'espace $S \subset \mathbb{L}_2([0,1])$ avec le produit scalaire usuel défini par $\langle f, g\rangle = \int_0^1 f(t)g(t)\,dt$, $f, g \in \mathbb{L}_2$ et $\|\cdot\|$ la norme associée. Dans l'expression précédente et le reste du document, $E$ désigne l'espérance sous la loi de $X$.

Supposons $E\left[\|X\|^2\right] < +\infty$, pour que (2) ait un sens et définissons

$$\begin{aligned}\Gamma : \mathbb{L}_2 &\to \mathbb{L}_2\\ f &\mapsto E[\langle X - E[X], f\rangle(X - E[X])],\end{aligned}$$

l'opérateur de covariance associé aux données. Comme $\Gamma$ est un opérateur compact et auto-adjoint, il existe une base hilbertienne de $\mathbb{L}_2$ composée de fonctions propres de $\Gamma$. D'autre part, il est possible de montrer que la solution de (2) est l'espace engendré par les fonctions propres $\psi_1^*, \ldots, \psi_D^*$ de l'opérateur $\Gamma$ associées aux $D$ plus grandes valeurs propres $\lambda_1^*, \ldots, \lambda_D^*$ (comptées avec multiplicité) et que cette solution est unique si ces valeurs propres sont toutes distinctes.

On va chercher à étudier l'erreur quadratique d'estimation des éléments propres en question à partir de données qui se présentent sous la forme discrétisée bruitée (1).

# 2 Nouveaux résultats pour la théorie minimax du problème

## 2.1 Le choix de modélisation de l'aléa

Nous adoptons une modélisation des variables aléatoires $X_i \overset{i.i.d.}{\sim} P_X$ avec $P_X \in \mathcal{P}$ une classe non-paramétrique de mesure de probabilité. En revanche, pour construire des estimateurs consistants, des contraintes sur la richesse de la classe $\mathcal{P}$ sont nécessaires. Nous définissons cette classe en deux temps :
On va s'intéresser à l'estimation des $\ell$-ièmes éléments propres de l'opérateur de covariance associé aux données, avec $\ell \in [\![1, p]\!]$.
On définit dans un premier temps une classe de noyaux dotés d'une certaine régularité pour tout $(m, L) \in {\mathbb{R}_+^*}^2$

$$\mathcal{K}(m,L) := \Bigg\{ K : [0,1]^2 \mapsto \mathbb{R} \;\; \lfloor m \rfloor\text{-differentiable} \Big| \, \forall u, u', t \in ]0,1[^3, K(u,t) = K(t,u),$$
$$\left| \frac{\partial^{\lfloor m \rfloor} K}{(\partial x_1)^{\lfloor m \rfloor}}(u,t) - \frac{\partial^{\lfloor m \rfloor} K}{(\partial x_1)^{\lfloor m \rfloor}}(u',t) \right| \leq L \left| u - u' \right|^{1+(m-\lfloor m \rfloor - 1)\mathbb{I}_{m \notin \mathbb{N}\setminus\{0\}}} \Bigg\}.$$

Puis on définit, sur l'espace des opérateurs compacts définis positifs de $\mathbb{L}_2([0,1])$ (qu'on notera $\mathcal{L}$), la forme suivante

$$\begin{aligned}\mathrm{r}_\ell : \mathcal{L} &\to \mathbb{R}\\ G &\mapsto r_\ell(G) := \frac{\lambda_\ell^*(G)\lambda_{\ell+1}^*(G)}{\left|\lambda_\ell^*(G) - \lambda_{\ell+1}^*(G)\right|^2} + \mathbb{I}_{\ell \neq 1} \frac{\lambda_\ell^*(G)\lambda_{\ell-1}^*(G)}{\left|\lambda_\ell^*(G) - \lambda_{\ell-1}^*(G)\right|^2},\end{aligned}$$

où $\lambda_j^*(G)$ correspond à la $j$-ième valeur propre de $G$.
Notons que des quantités analogues sont aussi apparues dans d'autres travaux liés à l'ACPf (voir la notion de "rang relatif" de Jirak et Wahl (2018) ou encore les travaux de Mas et Ruymgaart (2015)).

Enfin, on spécifie, pour toutes suites $(c_n)_{n\in\mathbb{N}\setminus\{0\}}, (d_p)_{p\in\mathbb{N}\setminus\{0\}} \in \mathbb{R}_+^{\mathbb{N}}$ telles que $c_n = o_n(n)$ et $d_p = o_p\left(p^m\right)$, la classe de mesures de probabilité modélisant la loi des $X_i$,

$$\mathcal{P}(\ell, c_n, d_p, m) := \Bigg\{ P, \text{ mesure de probabilité associée à un processus à trajectoires continues centrés et dont l'opérateur de covariance } \Gamma \text{ est tel que } \mathrm{r}_\ell(\Gamma) \leq c_n, \max\left(\frac{\lambda_1^*(\Gamma)}{\lambda_\ell^*(\Gamma)}, \lambda_\ell^*(\Gamma)^{-1}\right) \leq d_p, \left(\lambda_{\ell-1}^*(\Gamma)\lambda_\ell^*(\Gamma)\right)^2 = o_n(n) \text{ et que } \forall (s,t)\in[0,1]^2, K(s,t) := \int_{\mathcal{C}^0} z(t)z(s)\,dP(z) \in \mathcal{K}\left(m, 8(2\pi)^m \sum_{j=1}^{\ell-1} j^m \lambda_j^*(\Gamma)\right). \Bigg\},$$

on rappelle que $o_n(\cdot)$ et $o_p(\cdot)$ désignent des quantités négligeables face à la quantité $\cdot$ lorsque, respectivement, $n \to \infty$ et $p \to \infty$.

Nous justifierons ce choix de modèle de façon intuitive, puis nous verrons qu'il est soutenu théoriquement par des théorèmes d'inconsistance et empiriquement par des simulations numériques.

## 2.2 Les bornes inférieures minimax

Nous présenterons ce qui est à notre connaissance la première borne inférieure minimax d'estimation des fonctions propres qui prend en compte la discrétisation et l'impact des contraintes mises sur les spectres des opérateurs de covariance des processus en question. On peut aussi y ajouter une borne inférieure minimax d'estimation des valeurs propres pour obtenir le résultat suivant

**Théorème.** *(Borne inférieure minimax pour les $\ell$-ièmes éléments propres)*
Soit $\ell \in [\![1,p]\!], m \in \mathbb{R}_+^*$, $(c_n)_{n\in\mathbb{N}\setminus\{0\}}, (d_p)_{p\in\mathbb{N}\setminus\{0\}} \in \mathbb{R}_+^{\mathbb{N}}$ telles que $c_n = o_n(n)$ et $d_p = o_p\left(p^m\right)$.
Si on dispose de données sous la forme (1) avec $X_i \overset{i.i.d.}{\sim} P_X$, alors il existe $C, C' > 0$ qui ne dépendent pas des autres paramètres du modèle telles que

$$\inf_{\widehat{\psi}_\ell} \sup_{P_X\in\mathcal{P}(\ell,c_n,d_p,m)} \mathbb{E}\left[\left\|\widehat{\psi}_\ell - \psi_\ell^*\right\|^2\right] \geq C\left(\frac{c_n}{n} + \frac{d_p^2}{p^{2m}}\right),$$
$$\inf_{\widehat{\lambda}_\ell} \sup_{P_X\in\mathcal{P}(\ell,c_n,d_p,m)} \mathbb{E}\left[\left(\widehat{\lambda}_\ell - \lambda_\ell^*\right)^2\right] \geq C'\left(\frac{1}{n} + \frac{d_p^2}{p^{4m}}\right).$$

Dans les cas usuellement traités dans la littérature, on suppose simplement (en plus de la régularité du noyau de covariance) que nos processus sont tels qu'on a $a, C_0 \in \mathbb{R}_+^*$,

$$\begin{aligned} &\lambda_j(\Gamma) = C_0 j^{-(a+1)},\ \forall j,n,p \in \mathbb{N}\setminus\{0\}, \\ \text{ou } &\lambda_j(\Gamma) = C_0 \exp(-ja),\ \forall j,n,p \in \mathbb{N}\setminus\{0\}. \end{aligned} \tag{3}$$

Il s'agit d'un cas particulier de notre modèle. Le cas échéant donne à partir du Théorème le Corollaire suivant :

**Corollaire.** *(Borne inférieure minimax pour les $\ell$-ièmes éléments propres)*
Soit $\ell \in [\![1, p]\!], m \in \mathbb{R}_+^*$.
Sous le schéma d'observation (1) avec $X_i \overset{i.i.d.}{\sim} P_X$, en supposant qu'on a (3), il existe $C_\ell, C'_\ell > 0$, dont les seules dépendances en les paramètres du modèle sont en $\ell$, telles que

$$\inf_{\widehat{\psi}_\ell} \sup_{P_X \in \mathcal{P}(\ell, C_\ell, C_\ell, m)} \mathbb{E}\left[\left\|\widehat{\psi}_\ell - \psi_\ell^*\right\|^2\right] \geq C'_\ell \left(\frac{1}{n} + \frac{1}{p^{2m}}\right),$$
$$\inf_{\widehat{\lambda}_\ell} \sup_{P_X \in \mathcal{P}(\ell, C_\ell, C_\ell, m)} \mathbb{E}\left[\left(\widehat{\lambda}_\ell - \lambda_\ell^*\right)^2\right] \geq C'_\ell \left(\frac{1}{n} + \frac{1}{p^{4m}}\right).$$

## 2.3 Estimateurs minimax-optimaux

Dans un dernier temps, nous présenterons des estimateurs simples, basés sur une projection sur une base d'ondelettes, qui sont optimaux au sens minimax sous quelques hypothèses supplémentaires. Nous expliciterons le cheminement des idées ayant abouti à de tels estimateurs. Si on se place dans le cadre du corollaire on a alors le théorème suivant

**Théorème.** *(Borne supérieure pour les $\ell$-ièmes éléments propres)*
Soit $(\widehat{\lambda}_\ell, \widehat{\psi}_\ell)$ nos estimateurs.
On se place sous le schéma d'observation (1) avec $X_i \overset{i.i.d.}{\sim} P_X$, en supposant qu'on a (3), que les trajectoires des $X_i$ sont p.s. $m$-Hölder et qu'il existe $M > 0$ tel que $\|X\| \leq M$ p.s..
Alors il existe $C_{\ell,\sigma,M} > 0$ qui ne dépend que de $\ell, \sigma$ et $M$ telle que

$$\mathbb{E}\left[\left\|\widehat{\psi}_\ell - \psi_\ell^*\right\|^2\right] \leq C_{\ell,\sigma,M}\left(n^{-1} + p^{-2m}\right),$$
$$\mathbb{E}\left[\left(\widehat{\lambda}_\ell - \lambda_\ell^*\right)^2\right] \leq C_{\ell,\sigma,M}\left(n^{-1} + p^{-2m}\right).$$

**Remarque.** Les conditions du théorème précédent impliquent qu'il existe $C_\ell > 0$ telle que $P_X \in \mathcal{P}(\ell, C_\ell, C_\ell, m)$, et on se retrouve donc bien dans le même cadre que le Corollaire de borne inférieure.

De façon assez surprenante, et ce même si le problème est abordé de façon non-paramétrique avec des hypothèses de régularités, il n'y a en fait pas besoin d'étape de régularisation/lissage des données. Cette observation n'est pas spécifique à notre approche et semble être intrinsèque à l'aspect fonctionnel des données. On retrouve notamment des phénomènes et des bornes similaires pour un autre problème d'Analyse de Données Fonctionnelles défini sur le schéma d'observation (1) dans l'article de Cai et Yuan (2011).

# Bibliographie

Belhakem R., Picard F., Rivoirard, V. et Roche A. (2022) Minimax estimation of Functional Principal Components from noisy discretized functional data.

Cai, T.T. et Yuan, M. (2011), Optimal estimation of the mean function based on discretely sampled functional data : Phase transition, *The Annals of Statistics*, 39(5), pp. 2330–2355.

Ferraty F. et Vieu P. (2006), *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York. Theory and practice.

Hsing, T. et Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*.

Jirak M. et Wahl M. (2018), Relative perturbation bounds with applications to empirical covariance operators, *Advances in Mathematics*.

Mas A., Ruymgaart F. (2015), High Dimensional Principal Projections, *Complex Analysis and Operator Theory*, 9, 35-63.

Viviani R., Grön G. et Spitzer M. (2005) Functional principal component analysis of fmri data, *Human Brain Mapping*, 24.

Warmenhoven J., Bargary N., Liebl D., Harrison A., Robinson M.A., Gunning E., Hooker G. (2021), PCA of waveforms and functional PCA: A primer for biomechanics, *Journal of Biomechanics*, Volume 116.

# Curve registration for mechanistic models

Quentin Clairon[1] & John Fricks [2] & Mélanie Prague [1]

[1] *SISTM team, Université de Bordeaux, Inria Bordeaux Sud-Ouest, France, quentin.clairon@u-bordeaux.fr; melanie.prague@u-bordeaux.fr*
[2] *SISTM team, Université de Bordeaux, School of Mathematical and Statistical Sciences, Arizona State University, USA, jfricks@asu.edu*

**Résumé.** Nous nous intéressons au problème d'alignement de courbes dans une population de $n$ sujets ; c'est à dire à l'estimation de fonctions $\{h_i\}_{i=1,\dots,n}$ quantifiant des déformations temporelles altérant la dynamique des fonctions de regression $\{X_i\}_{i=1,\dots,n}$ à partir d'observations discrètes et bruitées des fonctions perturbées $\{X_i \circ h_i\}_{i=1,\dots,n}$. Dans notre cas, nous avons aussi accès à des informations a priori sur la dynamique des fonctions de regression originales représentées par des équations différentielles ordinaires (EDOs): $\dot{X}_i = f_{\xi_i}(X_i, t)$. Par ce mélange d'approche descriptive/non-paramétrique et causale/paramétrique, nous voulons inférer aussi précisément et exhaustivement que possible l'ensemble des effets d'un traitement sur une population donnée. Via le modèle causal, nous quantifions l'effet du traitement sur des mécanismes déjà identifiés et intégrés dans l'EDO sous la forme de covariables à estimer. A partir du modèle descriptif, nous inférons les effets plus généraux du traitement dus à des mécanismes ignorés par l'EDO mais pris en compte par les fonctions $\{h_i\}_{i=1,\dots,n}$. Le problème est formulé comme un modèle de régression non-linéaire dans un cadre mixte pour incorporer la variabilité inter-sujets. Nous confirmons ensuite sur données simulées la capacité de notre méthode à estimer l'effet d'un traitement à la fois sur l'évolution globale de certaines variables d'intérêt mais aussi sur des mécanismes spécifiques déjà identifiés. Nous concluons ce travail par l'analyse de données pré-cliniques et cliniques d'essais proposant des traitements contre le VIH.

**Mots-clés.** Alignement de courbes, modèles non linéaires à effets mixtes, modélisation mécaniste, essais cliniques.

**Abstract.** We tackle the curve registration problem in which we learn time-warping functions $\{h_i\}_{i=1,\dots,n}$ from noisy observations of registered curves $\{X_i \circ h_i\}_{i=1,\dots,n}$. Still, in our case a priori knowledge regarding the unregistered curve dynamics is available under the form of a parametric ordinary differential equations (ODE)s $\dot{X}_i = f_i(X_i, t)$. From this combination of descriptive non-parametric model and causal parametric one, we aim to locate as accurately and exhaustively as possible the effect of a given therapy on a treated population. From the causal representation, we quantify treatment effects on well identified mechanisms, specified as ODE parameter covariates. From the descriptive one, we infer global action of treatment due to other mechanisms missed by the ODE but accounted for by time-warping functions, leading to distorted dynamics for treated subjects compared to the control group. The joint estimation of time-warping functions and ODE parameters is then cast as a non-linear regression problem in a mixed effect setting to account for inter-subject variability. We then confirm on simulated data the capacity of our method to estimate treatment effects on the general evolution of some variables of interests as well as on specific mechanisms acting on the patient dynamic. We conclude this work by analyzing pre-clinical and clinical data from trials testing HIV cures.

**Keywords.** Curve registration, nonlinear mixed effect models, mechanistic modeling, clinical trials.

# 1 General Framing of the problem

In this work, we focus on the curve registration problem in the specific case where the dynamic of the unregistered curves are ruled by differential equations. We start by framing this problem within the classic curve registration setting: for a population of $n$ subjects, we aim to infer the time-warping functions $\{h_i\}_{i=1,...,n}$ distorting the original (or unregistered) regression functions $\{X_i\}_{i=1,...,n}$. For this, we have access for the $i$-th subject to $n_i$ noisy measurements of the distorted (registered) function $X_i \circ h_i$ given by:

$$y(t_{i,j}) = C\left(X_i(h_i(t_{i,j}))\right) + \epsilon_{i,j} \text{ with } \epsilon_{i,j} \sim N(0, \Sigma) \tag{1}$$

where $t_i := \{t_{i,j}\}_{j=1,...,n_i}$ and $\epsilon_{i,j} = \{\epsilon_{i,j}\}_{j=1,...,n_i}$ are respectively the $n_i$ measurement timepoints and the i.i.d centered measurement noise corrupting the observations. But here, we move a bit further from this classic framework by assuming that 1/the $X_i$ are $d$-dimensional state-variables and 2/possibly partially observed where $C : \mathbb{R}^d \longmapsto \mathbb{R}^{d_{obs}}$ is the observation function, with $d_{obs} \leq d$. That is, we consider that a relevant representation of each subject evolution requires a multiple outcome process even though some of its component cannot be directly measured. In order to learn the evolution of the unobserved variables, we assume that we have access to a priori knowledge regarding the interactions between the observed and unobserved part of $X_i$ given by an ordinary differential equation model (ODE)s:

$$\begin{cases} \dot{X}_{\xi_i} = f_{\xi_i}(X_{\xi_i}, t) \\ X_{\xi_i}(0) = x_{i,0}(\xi_i) \end{cases} \tag{2}$$

depending on the $d_\xi-$dimensional parameter $\xi_i$ and ruling the evolution of the unregistered function $X_i := X_{\xi_i}$.

Regarding $\xi_i$ parametrization, we adopt a nonlinear mixed effect (NLME) setting leading to the so-called NLME-ODE models:

$$\xi_{i,l} = \overline{\xi_l} + \sum_{m=1}^{M} \pi_{Gp_m}^{\xi_l} 1_{i \in Gp_m} + b_i^{\xi_l} \text{ for } l = 1, ...., d_\xi \tag{3}$$

with:

1. $\overline{\xi_l}$: the mean value for $\xi_{i,l}$ common to the whole population,
2. $\pi_{Gp_m}^{\xi_l}$ : the corrective term applied to the mean value $\overline{\xi_l}$ for subjects belonging to the group $Gp_m$ (with possibly $\pi_{Gp_m}^{\xi_l} = 0$),
3. $b_i^{\xi_l} \sim N(0, \sigma_{\xi_l}^2)$ : the subject-specific variations (with possibly $\sigma_{\xi_l}^2 = 0$).

The parametrization (3) makes it possible to aggregate information from the whole population to estimate $\overline{\xi} := \left(\overline{\xi_1}, ...., \overline{\xi_{d_\xi}}\right)$ as well as the regressors $\pi^\xi := \left(\pi_{Gp_m}^{\xi_1}, ....\pi_{Gp_m}^{\xi_{d_\xi}}\right)_{m=1,..,M}$ in a sparse data setting while allowing for inter-subject variability thanks to $b_i^\xi := \left(b_i^{\xi_1}, ....b_i^{\xi_{d_\xi}}\right) \sim N(0, \Sigma_{b^\xi} := diag(\sigma_{\xi_l}^2))$. We use a similar decomposition for the functions $\{h_i\}_{i=1,...,n}$ with a mean value $\overline{h}$ common to everyone, group specific distortion $\pi^h := \left(\pi_{Gp_1}^h, ....\pi_{Gp_M}^h\right)$ and random variations $b_i^h$. For inference purpose, we rely on the differential equation representation of $h_i$ proposed by Ramsay & Li [9]

to construct a global NLME-ODE model embedding both $\{\xi_i\}_{i=1,...,n}$ and $\{h_i\}_{i=1,...,n}$ as parameters. We then use the finite basis approximations $(\overline{h}(t,\beta_h), \pi^h(t,\beta_{\pi^h}), b_i^h(t,\beta_{b_i^h}))) \simeq \left(\overline{h}(t), \pi^h(t), b_i^h(t)\right)$ to consider the inference of $(\xi_i, h_i)$ as a classic estimation problem in a mixed effect framework.

# 2 Time-distorted-NLME-ODE models

As in Ramsay & Li [9], we assume that the time transformation functions can be represented as the solution of the second order ODEs $\ddot{h}_i = w_i\dot{h}_i$, equivalently reformulated as a first order one by:

$$\begin{cases} \dot{g}_i = w_i g_i \\ \dot{h}_i = g_i. \end{cases} \tag{4}$$

As required in curve registration setting, this parametrization is enough to ensure $h_i$ is monotonous on $[0,\, T]$ and strictly increasing if $\dot{h}_i(0) > 0$ without any conditions on $w_i$. This allows to turn to constrained inference problem of $h_i$ into the unconstrained one of $w_i$. From the original ODE (2), we can derive the equation followed by the warped function: $X_{h_i,\xi_i}(t) := X_{\xi_i}(h_i(t))$ by differentiation of a composed function:

$$\begin{cases} \dot{X}_{h_i,\xi_i}(t) = \dot{h}_i(t) f_{\xi_i}(X_{h_i,\xi_i}(t), h_i(t)) \\ X_{h_i,\xi_i}(0) = x_i(h_i(0)). \end{cases}$$

We add the constraint $h_i(0) = 0$ to enforce the absence of instantaneous temporal shift for all subjects. We then gather all equations into one to describe the evolution of the time-distorted NLME-ODE:

$$\begin{cases} \dot{X}_{h_i,\xi_i} = g_i f_{\xi_i}(X_{h_i\xi_i}, h_i) \\ \dot{g}_i = w_i g_i \\ \dot{h}_i = g_i \\ (X_{h_i,\xi_i}(0), g_i(0), h_i(0)) = (x_i(0), \dot{h}_i(0), 0). \end{cases} \tag{5}$$

The null function $w_i = 0$ leads to $h_i(t) = \dot{h}_i(0)t$ and so, setting $\dot{h}_i(0) = 1$ corresponds to an absence of instantaneous temporal distortion. So, the original ODE (2) is embedded into the time-distorted NLME-ODE models (5) as a particular case.

We are left with the choice of $w_i$ and $\dot{h}_i(0)$. For $w_i$, we choose the piecewise linear function:

$$w_i(t) = 2\sum_{k=0}^{K} \beta_{ik} \frac{t - t_k}{(t_{k+1} - t_k)^2} 1_{[t_k, t_{k+1}]}(t) \tag{6}$$

with uniformly reparted anchor points $\{t_k\}_{k=1,...K}$. We get $\beta_{ik} = \int_{t_k}^{t_{k+1}} w_i(s)ds$, so the parameters $\beta_i = (\beta_{i0,...}\beta_{iK})$ quantify the temporal distortion on each interval $[t_k, t_{k+1}]$. In particular, having $\beta_{ik} < 0$ indicates that the evolution speed of the i-th subject on $[t_k, t_{k+1}]$ is slowed down comparing to the dynamic implied by the original model (2).

Now, we account for group and subject specific variations with parametrization:

$$\begin{cases} \log \dot{h}_i(0) = \overline{\beta_0} + \sum_{m=1}^{M} \log \pi_{Gp_m}^{\beta_0} 1_{i \in Gp_m} + b_i^{\beta_0} \\ \beta_{ik} = \overline{\beta_k} + \sum_{m=1}^{M} \pi_{Gp_m}^{\beta_k} 1_{i \in Gp_m} + b_i^{\beta_k} \text{ for } k = 1, ...K \end{cases} \tag{7}$$

where

1. $\overline{\beta_k}$: the distortion common to the whole population,
2. $\pi_{Gp_m}^{\beta_k}$ : the additional ones applied to subjects belonging to $Gp_m$,
3. $b_i^{\beta_k} \sim N(0, \sigma_{\beta_k}^2)$ : the subject-specific variations.

We can reconstruct the mean function $\overline{w}$ or the group-specific one $\overline{w}_{Gp_m}$ for parametrization (7):

$$\begin{cases} \overline{w}(t) = 2\sum_{k=0}^{K} \overline{\beta_k} \frac{t-t_k}{(t_{k+1}-t_k)^2} 1_{[t_k,t_{k+1}]}(t) \\ \overline{w}_{Gp_m}(t) = 2\sum_{k=0}^{K} \left(\overline{\beta_k} + \pi_{Gp_m}^{\beta_k}\right) \frac{t-t_k}{(t_{k+1}-t_k)^2} 1_{[t_k,t_{k+1}]}(t) \end{cases}$$

and we can subsequently solve the ODE (4) for $w = \overline{w}$ (resp. $w = \overline{w}_{Gp_m}$) with initial conditions $\dot{\overline{h}}(0) = e^{\overline{\beta_0}}$ (resp. $\dot{\overline{h}}_{Gp_m}(0) = \pi_{Gp_m}^{\beta_0} e^{\overline{\beta_0}}$) to derive the corresponding time-warping function $\overline{h}$ (resp. $\overline{h}_{Gp_m}$).

The estimation of the time-distorted NLME-ODE (5) can be done in a same manner as the original model (2) with the addition of parameters $\overline{\beta} = (\overline{\beta_0}, ...\overline{\beta_K})$, $\pi^{\beta} = \left(\pi_{Gp_m}^{\beta_0}, ...., \pi_{Gp_m}^{\beta_K}\right)_{m=1,....M}$. Nonetheless, we aim to ensure that fidelity to the data is preferably explained by structural parameter variations internal to the assumed model rather than by external adjustment. For this, we add the prior distributions $\overline{\beta} \sim N(0, I_{K+1})$ and $\pi^{\beta} \sim N(0, I_{M(K+1)})$, this is equivalent to start with the null hypothesis that there is no time-warping. We estimate the population parameters $\left(\overline{\xi}, \pi^{\xi}, \Sigma_{b^{\xi}}, \overline{\beta}, \pi^{\beta}\right)$ with maximum a posteriori, the estimation uncertainty being quantified by Fisher Information matrix; all of the required numerical methods being implemented in Monolix [5].

# 3 Illustration

## 3.1 Numerical Experiment

We test on simulated data our ability to estimate the parameters of a time distorted NLME-ODE model, both the time-warping functions and the structural parameters. As tested model, we consider a time-warped and structurally identifiable version of the ODE proposed by [6] describing the evolution of target/infected CD4+T cells and HIV viral load:

$$\begin{aligned} \dot{T} &= 1 - \phi TV - d_T T \\ \dot{I} &= \phi TV - d_I I \\ \dot{V} &= \Lambda \overline{I} - cV. \end{aligned} \tag{8}$$

In this ODE, $T$ and $I$ respectively represent rescaled concentrations of target and infected cell populations and $V$, the viral load. The parameter $\phi$ is the viral infection rate of target cells, $\Lambda$ the viral production rate and the remaining terms $d_T, d_I$ and $c$ are respectively the death rate of $T$ , $I$ and the viral clearance rate. The model is used to predict viral load rebound just after the interruption of a long antiretroviral therapy (ART) exposure, so we assume at $t = 0$ there is no infection ($\phi \simeq 0$) and the system is near its steady state $(T(0), I(0), V(0)) = (1/d_T, I_0, 0)$ but with $I_0 \neq 0$ to mimic the existence of viral reservoir in long-lived lymphocytes [10].

We use the mean parameter value given in Prague et al. [8, 7], $\phi = 7.7e^{-7}$, $d_T = 0.05$, $d_I = 0.4$, $c = 23$ as well as $\Lambda = 2.5e^6$ and $I(0) = 2e^{-6}$. Regarding the applied temporal distortion to ODE (8),

| Parameter | Target Value | Mean Estimates | Relative Bias | Empirical Std | Estimated Std | MSE | Coverage Rate |
|---|---|---|---|---|---|---|---|
| $\overline{\Lambda}$ | $2.5e^6$ | $2.6e^6$ | 0.02 | $1.3e^5$ | $1.8e^5$ | $1.6e^{10}$ | 0.93 |
| $\pi^{\overline{\Lambda}}_{Gp_1}$ | 0.8 | 0.8 | 0.01 | $5.3e^{-2}$ | $4.0e^{-2}$ | $2.9e^{-3}$ | 0.91 |
| $\pi^{\overline{\Lambda}}_{Gp_2}$ | 1.0 | 1.0 | 0.01 | $6.6e^{-2}$ | $4.9e^{-2}$ | $4.5e^{-3}$ | 0.90 |
| $\overline{I}(0)$ | $2.0e^{-6}$ | $1.9e^{-6}$ | <0.01 | $1.8e^{-7}$ | $1.7e^{-7}$ | $2.5e^{-14}$ | 0.95 |
| $\pi^{\beta_1}_{Gp_1}$ | -0.7 | -0.7 | 0.02 | $8.2e^{-2}$ | $7.5e^{-2}$ | $6.8e^{-3}$ | 0.93 |
| $\pi^{\beta_2}_{Gp_1}$ | -1.9 | -1.9 | <0.01 | $7.6e^{-2}$ | $7.5e^{-2}$ | $5.8e^{-3}$ | 0.93 |
| $\pi^{\beta_1}_{Gp_2}$ | -1.5 | -1.5 | <0.01 | 0.1 | 0.1 | $9.0e^{-3}$ | 0.97 |
| $\pi^{\beta_2}_{Gp_2}$ | -2.3 | -2.3 | 0.01 | 0.2 | 0.2 | $3.3e^{-2}$ | 0.91 |

Table 1: Estimation results for time-distorted ODE (8)

we resume the finite basis decomposition (6) for $w_i$ with $K = 2$ elements and $t_{k+1} - t_k = 30$ days between anchor points but we do not add mean temporal distortion by setting $\overline{\beta_0} = \overline{\beta_1} = \overline{\beta_2} = 0$.

We consider $M = 2$ treatment groups in addition to a placebo one. To account for both group and patient specific variations, we use the following parametrization:

$$\begin{cases} \ln \Lambda_i &= \ln \overline{\Lambda} + \ln \pi^{\overline{\Lambda}}_{Gp_1} 1_{i \in Gp_1} + \ln \pi^{\overline{\Lambda}}_{Gp_2} 1_{i \in Gp_2} + b_i^{\Lambda} \\ \ln I_i(0) &= \ln \overline{I}(0) + b_i^{I(0)} \\ \beta_{i0} &= b_i^{\beta_0} \\ \beta_{i1} &= \pi^{\beta_1}_{Gp_1} 1_{i \in Gp_1} + \pi^{\beta_1}_{Gp_2} 1_{i \in Gp_2} + b_i^{\beta_1} \\ \beta_{i2} &= \pi^{\beta_2}_{Gp_1} 1_{i \in Gp_1} + \pi^{\beta_2}_{Gp_2} 1_{i \in Gp_2} + b_i^{\beta_2} \end{cases}$$

with $\sigma^2_{\Lambda} = 0.3$ and $\sigma^2_{I(0)} = 1.0$, $\sigma^2_{\beta_0} = \sigma^2_{\beta_1} = \sigma^2_{\beta_2} = 0.2$ as well as regressors $\left(\pi^{\overline{\Lambda}}_{Gp_1}, \pi^{\beta_1}_{Gp_1} \pi^{\beta_2}_{Gp_1}\right) = (0.8, -0.7, -1.9)$ and $\left(\pi^{\overline{\Lambda}}_{Gp_2}, \pi^{\beta_1}_{Gp_2}, \pi^{\beta_2}_{Gp_2}\right) = (1.0, -1.5, -2.3)$.

We simulate a population of $n = 150$ subjects uniformly reparted within each group $m = 0, ..., M$ ($m = 0$ representing the placebo one). We generate for each subject $n_i = 21$ viral load measurements given by $Y_{i,j} = \log_{10}(V(t_{i,j})) + \varepsilon_{i,j}$ where $\varepsilon_{i,j} \sim N(0, 0.2^2)$ with two measurements a week for the first four weeks followed by one measurement each week afterward up to 40 weeks. From this synthetic dataset, we estimate the mean parameter value $\left(\overline{\Lambda}, \overline{I}(0)\right)$ as well as the regressors $\left(\pi^{\overline{\Lambda}}_{Gp_1}, \pi^{\beta_1}_{Gp_1} \pi^{\beta_2}_{Gp_1}\right)$ and $\left(\pi^{\overline{\Lambda}}_{Gp_2}, \pi^{\beta_1}_{Gp_2}, \pi^{\beta_2}_{Gp_2}\right)$.

We proceed to $N_{MC} = 10^3$ trials of such data simulation and subsequent parameter estimation. From these trials, we estimate the mean, the bias, the empirical standard deviation as well as the mean square error of our estimator to quantify its practical accuracy. We also estimate the standard deviation derived from the Fisher Information Matrix as well as the coverage rate of the corresponding 95% confidence intervals. The estimation results are given in table 1. We can see that both empirical and estimated standard deviation globally coincides and the actual coverage rate is generally close to the expected rate of 0.95, thus indicating a well-conditioned estimation problem. Still, we denote a slight under-estimation of variance for regressors linked to $\Lambda$ consistent with the small drop in coverage rate for $\pi^{\overline{\Lambda}}_{Gp_1}$ and $\pi^{\overline{\Lambda}}_{Gp_2}$.

## 3.2 Real Data

We will apply this methodology to trials in which HIV cure therapies and vaccines have been tested either in non-human primates [1, 4, 8] and humans [3, 2]. The dataset consists in Antiretroviral treatment interruption (ATI) trials, which are pivotal in the landscape of HIV cure research, offering a structured framework to evaluate the efficacy of cure strategies by temporarily halting antiretroviral therapy (ART) under stringent medical oversight. These trials aim to elucidate the immune system's capacity to control HIV in the absence of medication, thereby shedding light on potential pathways to achieve viral remission or cure. In situations where effective or functional cures are not available (as for nowadays), the impact of any intervention is likely to result from a combination of intrinsic alterations in the virus system's mechanistic capabilities, along with a delayed response that occurs at varying rates across individuals. The issue of this variability, particularly the delayed response that differs from one person to another, is addressed through the technique of curve registration.

# Bibliographie

# References


[1] Erica N Borducchi, Crystal Cabral, Kathryn E Stephenson, Jinyan Liu, Peter Abbink, David Nganga, Joseph P Nkolola, Amanda L Brinkman, Lauren Peter, Benjamin C Lee, et al. Ad26/mva therapeutic vaccination with tlr7 stimulation in siv-infected rhesus monkeys. *Nature*, 540(7632):284–287, 2016.

[2] Amanda Cobb, Lee K Roberts, A Karolina Palucka, Holly Mead, Monica Montes, Rajaram Ranganathan, Susan Burkeholder, Jennifer P Finholt, Derek Blankenship, Bryan King, et al. Development of a hiv-1 lipopeptide antigen pulsed therapeutic dendritic cell vaccine. *Journal of immunological methods*, 365(1-2):27–37, 2011.

[3] Y Lévy, C Lacabaratz, Edouard Lhomme, A Wiedemann, Claire Bauduin, C Fenwick, E Foucat, M Surenaud, L Guillaumat, Valerie Boilet, et al. A randomized placebo-controlled efficacy study of a prime boost therapeutic vaccination strategy in hiv-1-infected individuals: Vri02 anrs 149 light phase ii trial. *Journal of Virology*, 95(9):10–1128, 2021.

[4] So-Yon Lim, Christa E Osuna, Peter T Hraber, Joe Hesselgesser, Jeffrey M Gerold, Tiffany L Barnes, Srisowmya Sanisetty, Michael S Seaman, Mark G Lewis, Romas Geleziunas, et al. Tlr7 agonists induce transient viremia and reduce the viral reservoir in siv-infected rhesus macaques on antiretroviral therapy. *Science translational medicine*, 10(439):eaao4521, 2018.

[5] Lixoft. Monolix version 2023r1. *Antony, France*, 2023.

[6] Alan S Perelson and Ruy M Ribeiro. Modeling the within-host dynamics of hiv infection. *BMC biology*, 11:1–10, 2013.

[7] Mélanie Prague, Daniel Commenges, Julia Drylewicz, and Rodolphe Thiébaut. Treatment monitoring of hiv-infected patients based on mechanistic models. *Biometrics*, 68(3):902–911, 2012.

[8] Mélanie Prague, Jeffrey M Gerold, Irene Balelli, Chloé Pasin, Jonathan Z Li, Dan H Barouch, James B Whitney, and Alison L Hill. Viral rebound kinetics following single and combination immunotherapy for hiv/siv. *BioRxiv*, page 700401, 2019.

[9] James O Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(2):351–363, 1998.

[10] Janet D Siliciano and Robert F Siliciano. The latent reservoir for hiv-1 in resting cd4+ t cells: a barrier to cure. *Current Opinion in HIV and AIDS*, 1(2):121–128, 2006.

# Bayesian Registration using Hamiltonian Monte Carlo.

Henrique Cheng [1] & John Fricks [2]

[1] *School of Mathematical and Statistical Sciences, Arizona State University, USA hcheng43@asu.edu*
[2] *School of Mathematical and Statistical Sciences, Arizona State University, USA & SISTM team, Bordeaux Public Health, Université de Bordeaux, France jfricks@asu.edu*

**Résumé.** Dans cette présentation, une nouvelle méthode d'enregistrement de courbe dans le contexte de l'estimation de la fonction moyenne est présenté. Un schéma numérique Hamiltonien de Monte Carlo est utilisé pour échantillonner la distribution a posteriori des fonctions de déformation étant donné les données fonctionnelles. Une comparaison numérique avec un algorithme de Metropolis-Hastings plus traditionnel est présentée. La conclusion préliminaire de l'étude numŕique est que l'approche informatique Hamiltonienne de Monte Carlo accélère considérablement la convergence vers le postérieur de haute dimension requis pour l'estimation de la fonction de courbe.

**Mots-clés.** Données fonctionnelles, Enregistrement de courbe, Statistique bayésienne.

**Abstract.** In this presentation, a novel method for curve registration in the context of mean function estimation is presented. A Hamiltonian Monte Carlo numerical scheme is used to sample for the posterior distribution of the warping functions given the functional data. A numerical comparison with a more traditional Metropolis-Hastings algorithm is presented. The preliminary conclusion of the numerical study is that the Hamiltonian Monte Carlo computational approach substantially speeds convergence to the high dimensional posterior that is required for curve function estimation.

**Keywords.** Functional Data Analysis, Curve Registration, Bayesian Statistics.

## 1 Introduction

Registration is an important topic of functional data analysis. Intuitively, registration attempts to remove phase variation in a sample of observed functions. For an extended discussion on registration see Ramsay and Silverman (2019) or Marron et al (2015). One goal of registration can be to infer the mean function of such a group of functions in the presence of phase variation. We will largely be following Earls and Hookers approach to simultaneously register and estimate a mean function (2017).

We write a more specific statistical model as follows.

$$Y_j\left(t_{i,j}\right) = \mu\left(h_j\left(t_{i,j}\right)\right) + X_j\left(h_j\left(t_{i,j}\right)\right) + \epsilon_{i,j} \tag{1}$$

for time index $i = 1, ..., I_j$, subject index $j = 1, ..., J$, and with $0 \leq t_{i,j} \leq T$. The goal for the purposes of this study, then, will be to estimate $\mu(\cdot)$.

The function $h_j(\cdot)$ is a subject-specific increasing time warping function with $h_j(0) = 0$ and which will be represented here as

$$h_j(t) = \int_0^t e^{w_j(s)} ds \tag{2}$$

with $w_j(\cdot)$ modeled as a zero-mean Gaussian, Matérn process.

# 2 Simulation Setup and Conclusion

A Bayesian registration method was implemented using a Hamiltonian Monte Carlo method for sampling from the posterior distribution of $h_j|Y_j$ following Betancourt (2017). For testing purposes, a mean model consisting of a finite combination of Fourier components with Gaussian amplitude error and a time-warping function described above. The full hierarchical Bayes model is given here:

$$\begin{aligned}
Y_j(h_j(t))|f(t|\boldsymbol{\beta}, \tau) &\overset{iid}{\sim} GP(\mu(t), \Sigma_f(t, u|\sigma_f^2)); \quad t, u \in \mathcal{T}, \quad j = 1, ..., J \\
h_j(t)|w_j(t) \qquad & w_j(t) \overset{iid}{\sim} GP(0, \Sigma_w(t, u|\boldsymbol{\gamma})) \\
& \boldsymbol{\gamma} = (\gamma_1, \gamma_2)' \\
\mu(t|\boldsymbol{\beta}, \tau) &= \beta_0 + \sum_{q=1}^{Q} \beta_{1q} \cos(\frac{2\pi q}{\tau} t) + \beta_{2q} \sin(\frac{2\pi q}{\tau} t) \\
& \boldsymbol{\beta} \sim \mathrm{N}_{2Q+1}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\
& \tau \sim N(\mu_\tau, \sigma_\tau^2) \\
& \sigma_f^2 \sim IG(a/2, 2/b) \\
& \boldsymbol{\gamma} \sim N_2(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}})
\end{aligned}$$

For comparison, a Metropolis-Hastings within Gibbs Monte Carlo method for sampling the posterior distribution was also developed. Here we present some preliminary simulation results. In Figure 1, we see three realizations simulated from the model in Equation 1 with

$$\mu(t) = -1 \cos(\frac{2\pi}{12} t) + 2 \cos(\frac{2\pi}{6} t) - 2 \cos(\frac{2\pi}{4} t)$$

and $\sigma_\epsilon^2 = 0.55^2$, as well as with the warping function defined by Equation 2 with $w(\cdot)$ being a Matérn process with $(\sigma_w^2, \rho, \nu)' = (0.2, 1, 3)$.

These three simulations were used as data and the parameters of the Bayesian hierarchical model presented above were estimated by using the Hamiltonian Monte Carlo method to draw from the posterior distributions. A more traditional Metropolis within Gibbs was also used for comparison. Posterior means can be found in Table 1.

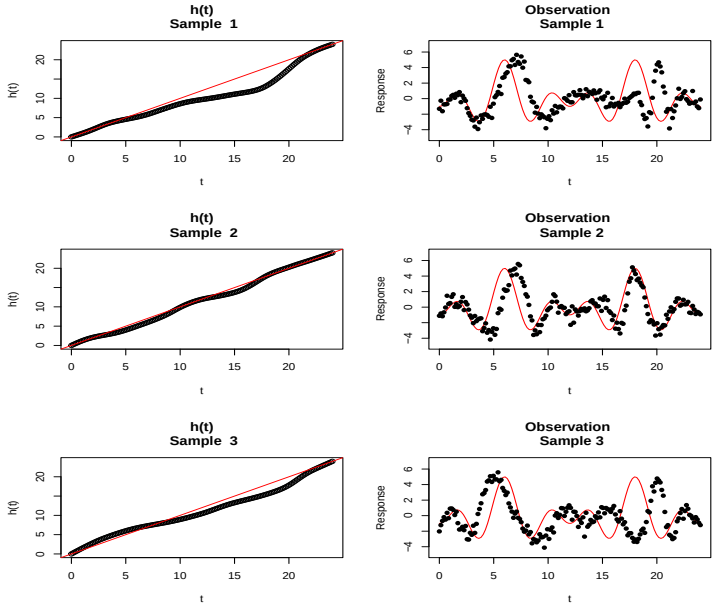



Figure 1: Three simulated warping functions whose base values are independently sampled from a MVN distribution with Matérn covariance function parameters $(\sigma_w^2, \rho, \nu)' = (0.2, 1, 3)$ on the LHS. On the RHS, each respective warping function is applied to 170 points along the reference function $-1\cos(\frac{2\pi}{12}t) + 2\cos(\frac{2\pi}{6}t) - 2\cos(\frac{2\pi}{4}t)$ (red line) and iid $N(0, .55^2)$ amplitude noise is added to generate observations (black points).

| S1 | $\tau$ | $\boldsymbol{\beta}$ | $\sigma_f^2$ | $\exp(\boldsymbol{\gamma})$ |
|---|---|---|---|---|
| True | 12 | $(0, -1, 0, 2, 0, -2, 0)$ | $0.55^2$ | $(0.2, 1)$ |
| MH | 12.69 | $(-0.09, -0.57, 0.07, 0.94, 0.23, -0.51, -0.17)$ | $1.87^2$ | $(0.43, 0.72)$ |
| HMC | 12.08 | $(-0.05, -0.93, -0.03, 1.98, -.01, -1.93, .01)$ | $0.53^2$ | $(0.28, 0.91)$ |

Table 1: A comparison of true parameter values and point estimates under the strict Metropolis within Gibbs and HMC adapted sampler.

In addition to the point estimates, the posterior warping functions can be sampled. In Figure 2, the simulated as well as the posterior warping function for both the HMC and Metropolis within Gibbs samplers for each of the three replicates. Note that the posterior means of the HMC are a substantially closer fit to the true warping function compared to the Metropolis within Gibbs sampler for comparable computational cost.

To see the effect of registration on the fit of the reference function, Figure 3 shows simulated data along with the registered reference functions. The last column shows the comparison using the HMC algorithm with the Metropolis within Gibbs shown in column 2. The plots show substantial improvement of the fit using the HMC algorithm with comparable computing cost, bolstered by the root mean squared error of roughly 0.5 for HMC with root means squared error for the Metropolis within Gibbs close to 2.

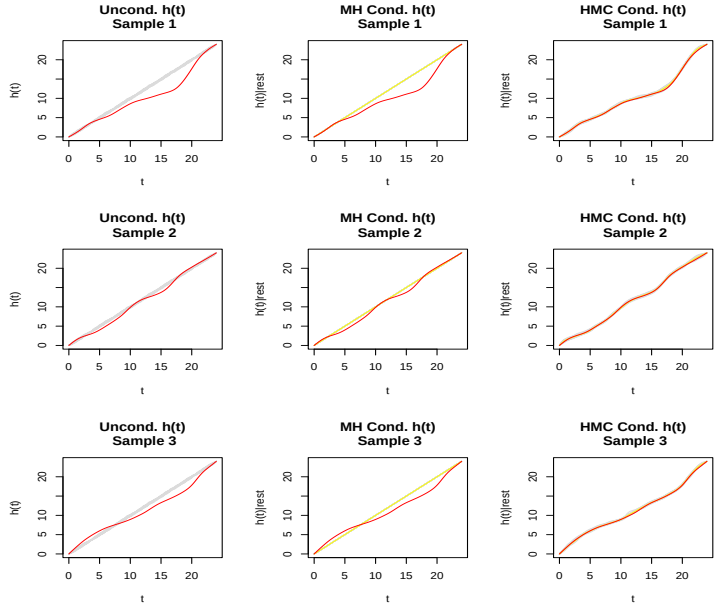



Figure 2: Linearly interpolated draws from the unconditional (first column) and conditional (second and third columns) warping functions for each sample in the study. The true warping function is in red for each plot, the yellow lines in the second column are the estimated warping functions under the strict Metropolis within Gibbs sampler for each sample, and the yellow lines in the third column are likewise but for the HMC adapted sampler. In the second column, each $8000^{th}$ draw was taken from the posterior distribution for each warping function and in the third column, each $2400^{th}$ draw was taken.

The primary conclusion from the numerical results is that similar results can be obtained from the two methods; however, the more traditional Metropolis-Hastings within Gibbs sampler required substantially more computational resources to accomplish a similar statistical accuracy to HMC.

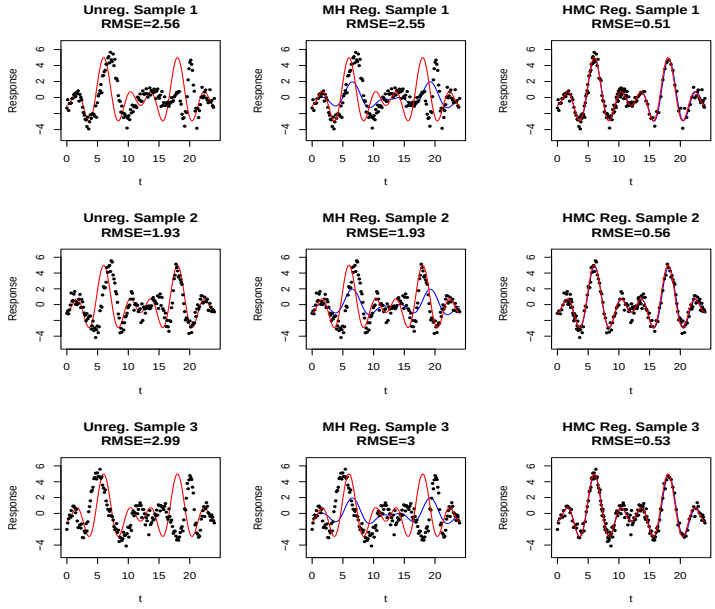



Figure 3: Points corresponding to the unregistered (left-most column) and registered values under the strict Metropolis within Gibbs and HMC adapted samplers (center and right-most columns) for the study. The red line corresponds to the true reference function, whereas the blue lines in the second and third columns correspond to the estimated reference function under the Metropolis and HMC samplers, respectively. The RMSE values were computed by taking the square root of the average squared deviances of response values from the true reference function.

# Bibliography

Marron, James Stephen, James O. Ramsay, Laura M. Sangalli, and Anuj Srivastava. "Functional data analysis of amplitude and phase variation." Statistical Science (2015): 468-484.

Ramsay, J. and B. Silverman. (2005). Functional data analysis (Second ed.). New York: Springer.

Earls, C. and G. Hooker. (2017). Variational Bayes for functional data registration, smoothing, and prediction. Bayesian Analysis 12(2): 557-582.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434.

# Landmark and elastic registration of aircraft trajectories

Rémi Perrichon [1] & Xavier Gendre [2,3] & Thierry Klein [1,2]

[1] *École Nationale de l'Aviation Civile, Université de Toulouse, Toulouse, France.*
*remi.perrichon@enac.fr*
*thierry01.klein@enac.fr*
[2] *Institut de Mathématiques de Toulouse (UMR5219), Université de Toulouse, Toulouse, France.*
*xavier.gendre@math.univ-toulouse.fr*
[3] *Pathway.com, Paris, France.*

**Résumé.** L'étude statistique de trajectoires d'avions dans le cadre de l'analyse des données fonctionnelles nécessite un ensemble de pré-traitements bien connus de la littérature. On s'intéresse ici au problème d'alignement de données de trajectoires, fréquent dans de nombreuses applications. Pour les données de trajectoires, la présence de variations de phase est généralement inévitable car les vols étudiés ne se déroulent jamais simultanément, ont des durées différentes et sont opérés avec de grandes variabilités selon les compagnies aériennes. Nous proposons une comparaison de deux méthodes d'alignement : un alignement dit "par landmarks" et un alignement dit "élastique". A partir d'un échantillon de vols, l'objectif est de constituer un profil moyen d'altitude. Ce profil moyen doit être le plus informatif possible au sens où on souhaite qu'il caractérise les amplitudes moyennes de l'altitude. Une bonne procédure d'alignement devrait donc permettre de résumer les variations de l'altitude pour des phases de vol similaires. Dans le cas idéal où les phases de vol sont renseignées dans les données brutes, l'alignement par landmarks est naturel : il suffit d'identifier les landmarks aux changements de phase. Si ce n'est pas le cas, on peut préalablement segmenter les phases de vol. Moyennant un cadre conceptuel plus avancé reposant sur la géométrie différentielle, nous montrons que l'alignement élastique produit un profil d'altitude moyen plus pertinent. Grâce à une métrique riemannienne bien choisie, l'alignement permet de bien distinguer les plateaux de la phase d'approche, alors même que l'information sur les phases de vol n'est pas explicitement utilisée.

**Mots-clés.** Analyse de données fonctionnelles, alignement, métrique élastique, trajectoires.

**Abstract.** The statistical analysis of aircraft trajectories within the framework of Functional Data Analysis (FDA) requires a set of preprocessing steps that are well known in the literature. We are interested in the problem of aligning trajectory data, which is common in many applications. For trajectory data, the presence of phase variations is generally inevitable because the studied flights are never simultaneous, have different durations, and are operated with significant variabilities across airlines. We propose a comparison of two registration methods: a landmark-based registration and an elastic registration. From a sample of flights, the objective is to construct an average altitude profile. This average profile should

be as informative as possible in the sense that it should characterize the average altitude amplitude. A good registration procedure should therefore summarize altitude variations for similar flight phases. In the ideal scenario where flight phases are provided in the raw data, landmark registration is natural: it suffices to identify landmarks at phase changes. If this is not the case, flight phases can be segmented beforehand. Leveraging a more advanced conceptual framework based on differential geometry, we highlight that elastic registration produces a more relevant average altitude profile. Thanks to a well-chosen Riemannian metric, the registration enables the clear distinction of plateaus in the approach phase, even when information about flight phases is not explicitly used.

**Keywords.** Functional data analysis, registration, elastic metric, trajectories.

# 1 Introduction

Thanks to advances in data storage, the growing use of sensors and the development of advanced computational techniques, it is not uncommon for statisticians to manipulate statistical units such as images, sounds, or curves. These new objects have prompted the emergence of a new terminology in statistics. The phrase Object Oriented Data Analysis (OODA), was defined by Wang and Marron (2007) to be "the statistical analysis of populations of complex objects". Part of OODA is Functional Data Analysis (FDA) for which the atoms of the statistical analysis are functions. The term FDA was coined by Ramsay (1982) and Ramsay and Dalzell (1991), even though the origin of FDA can be traced back much earlier as explained by Müller (2016). The foundational monograph of Ramsay and Silverman (2005), the work of Kokoszka and Reimherr (2021) and the review of FDA techniques written by Wang, Chiou, and Müller (2016) may serve as good introductions to the topic.

*Second-generation functional data* have recently been defined by Koner and Staicu (2023) as "functional data acquired in a multivariate, longitudinal, time series, or spatial design". Multivariate functional data are typically defined on the unit interval as vector-valued functions in $\mathbb{R}^d$ ($d > 1$). Aircraft trajectories have traditionally been studied as such. To be precise, they have often been modeled as parametric paths in the sense of differential geometry, as originally proposed by Puechmorel and Delahaye (2007) and Delahaye et al. (2014). It resulted in several promising applications. Based on a sample of aircraft landing trajectories at Toulouse-Blagnac airport, Suyundykov, Puechmorel, and Ferré (2010) have identified major flows around an airport. A detection of bad runway conditions has been developed by Andrieu et al. (2016).

When considering flights in their entirety, that is to say from takeoff to landing, it is necessary to register trajectories before any statistical analysis. Registration is a standard pre-processing step in FDA that aims at separating phase variations from amplitude variations. Without this preliminary transformation, Marron et al. (2015) highlighted that a statistical analysis as basic as averaging may not offer an effective data summary. For example, the shifted betas example developed by Marron and Dryden (2021) (Section 9.1, p.176)

illustrates the limitations of Functional Principal Component Analysis (FPCA) in the presence of phase variations. These same limitations have been identified for observational data, notably by Nicol (2013), who demonstrated the importance of registration for the Functional Principal Component Analysis (FPCA) of aircraft trajectories.

Phase variations are inevitable as flights are neither simultaneous nor of the same duration and exhibit strong operational variations. The key is to find an effective method to compare them at similar time points, meaning, literally, for the same flight phases. In this work, we compare two commonly used approaches in the alignment of multivariate functional data: a landmark-based approach and an elastic registration approach. The goal is to construct an effective data summary of the average altitude profile. We extend the discussion initiated by Marron et al. (2014) and companion papers to the case of commercial aviation. We show that even in the ideal scenario where the landmark approach fully takes advantage of a known segmentation of flight phases, elastic registration yields a more comprehensible average altitude profile provided a wise choice of component.

In the following, we consider a sample of $n = 5$ flights over the United States made available by the National Aeronautics and Space Administration (NASA). Each flight is observed at a finite number of moments, that is we observe

$$(t_{ij}, \boldsymbol{y}_{ij}),\ t_{ij} \in [0,1],\ \boldsymbol{y}_{ij} \in \mathbb{R}^d,\ i = 1, ..., n,\ j = 1, ..., J_i \tag{1}$$

where $\boldsymbol{y}_{ij} \equiv (y_{ij}^{[1]}, ..., y_{ij}^{[d]})$ $(d > 1)$. For a given flight $i$, $J_i$ values are observed for all $d$ components. Typically, the first three components describe the position of the aircraft (longitude, latitude, altitude). Speed, acceleration and weather values are classic examples of the other components. Since we are dealing with domestic flights, we do not consider the angular nature of longitude and latitude in this work. Time has been scaled such that the first point of each trajectory is associated with $t = 0$ and the last point with $t = 1$. The high sampling rate enables individual smoothing of each trajectory.

# 2 A landmark approach to register aircraft trajectories

Originally introduced by Kneip and Gasser (1992) and Gasser and Kneip (1995), landmark registration is a popular approach to align functional data. It easily adapts to the case of multivariate functional data and relies only on a handful of steps. For each flight, some structural features must first be identified as well as their timings. A template is then chosen (often, the mean of the timings). A strictly monotonic time-warping function is computed so that, for each trajectory, landmarks occur at the same time as in the template. Finally, registered values are obtained using the inverse warping function. All details may be found in Ramsay and Silverman (2005) (Section 7.3, p.132). The alignment's accuracy depends on clearly defining the features. Two cases arise in aviation depending on whether the flight phases are already labeled in the raw data or not.

## 2.1 Registration when flight phases are labeled in the raw data

When flight phases are already identified in the raw data, the structural features naturally correspond to the beginning (or end) of each flight phase. Their timings are explicitly available, making this situation an ideal scenario. In practice, it is the case for Flight Data Recorder (FDR) data because flight phases are automatically determined based on the monitoring and recording of many flight parameters. These are the data we use. The chosen landmarks are the takeoff, the last point of the climb phase, the first and last point of the approach. Figure 1 illustrates the impact of landmark registration on the determination of an average altitude profile. Note that the use of monotone cubic Hermite spline interpolation instead of linear interpolation for constructing the time warping functions seems that have a little effect on the obtained average altitude profile. In either case, the pronounced plateaus of the approach phase observed in trajectories n°1 and n°4 are not reflected in the average altitude profile.

## 2.2 Registration when flight phases are not labeled in the raw data

In the vast majority of cases, flight phases are not labeled in raw data. Several approaches are thus possible. First, we can select features based on peaks, points of inflection, and threshold crossings of one or more components of the trajectory and/or their derivatives. We then hope to *implicitly* retrieve the different flight phases and apply the registration steps as usual. A second approach consists of *explicitly* identifying flight phases using algorithms present in the literature of aviation transportation. Note that segmenting a flight into different phases is generally a complex task. Indeed, there are substantial operational differences attributable to weather conditions and/or air traffic control. Even within the same phase, aircraft may climb at different rates or fly at different cruise altitudes. As a consequence, specifying universal thresholds for flight phase segmentation is not the most effective approach. Commonly used approaches in the literature have recently been reviewed by Fala, Georgalis, and Arzamani (2023). Among statistical methods, Perrichon, Gendre, and Klein (2024) have recently demonstrated the usefulness of using Hidden Markov Models (HMMs) to identify the main flight phases of commercial aviation with high precision.

In all cases, landmark registration is only discrete evidence concerning intrinsically continuous warping functions. It ignores what happens in between landmarks, which is why it is customary to adopt a continuous fitting criterion for registration. In our case, it would be desirable to identify certain prominent plateaus.

# 3 Elastic registration of aircraft trajectories

Elastic registration relies on a continuous fitting criterion. It has been developed by Anuj Srivastava et al. (2011) to tackle several limitations of the $\mathbb{L}^2$ distance registration. It provides a natural template for the alignment and allows for rigorously defining the concepts of phase and amplitude variations. It is based on differential geometry. In the same perspective as

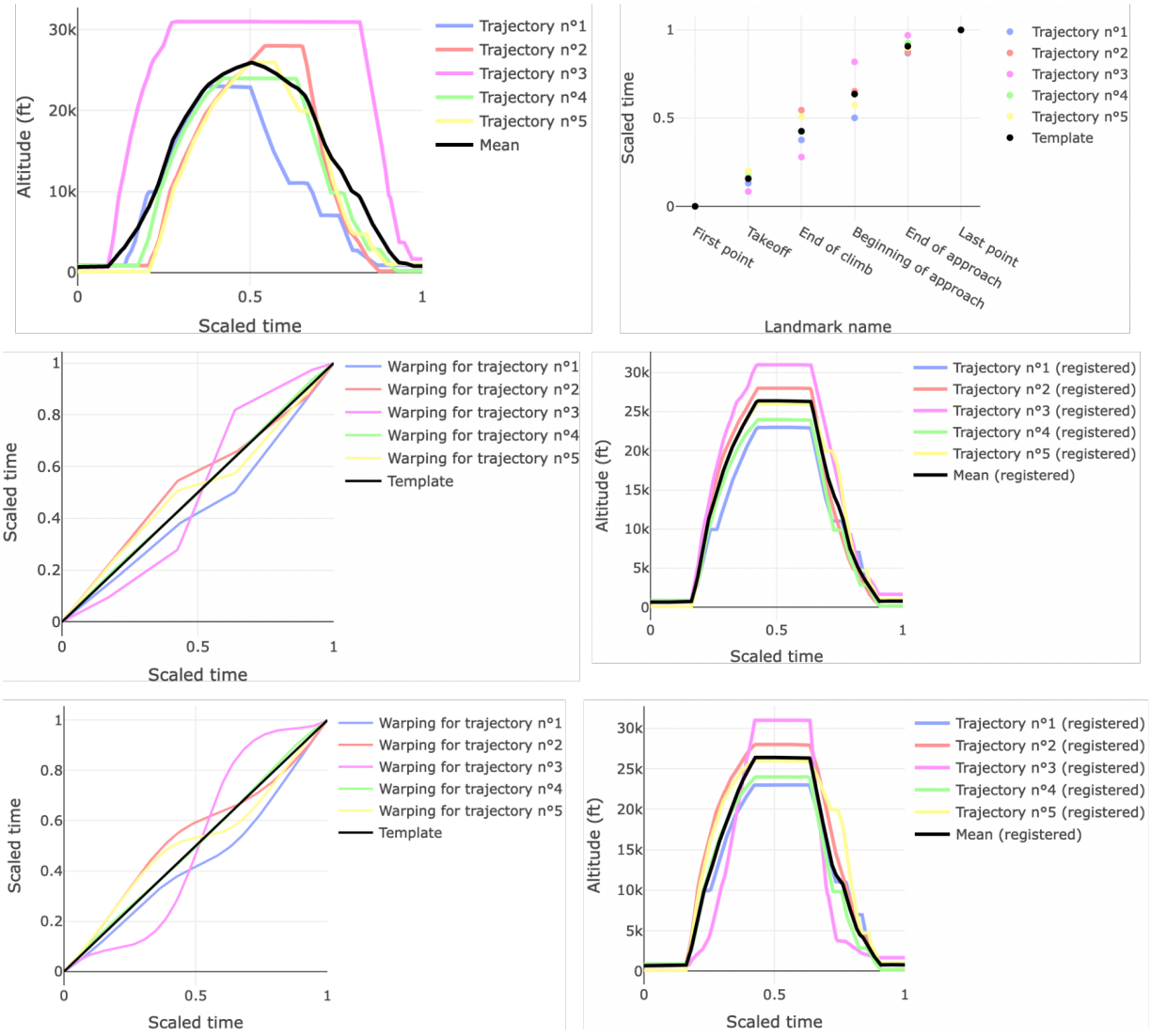



Figure 1: Altitude profiles and empirical average for raw data [top left], identification of landmarks, their timings, and a template based on the average [top right], calculation of time warping functions using linear interpolation [middle left], registered altitude profiles and the obtained registered empirical average when warping functions have been constructed with linear interpolation [middle right], time warping functions using monotone cubic Hermite spline interpolation [bottom left], registered altitude profiles and the obtained registered empirical average when warping functions have been constructed with monotone cubic Hermite spline interpolation [bottom right].

for statistical shape analysis, the foundational idea is to consider a relevant quotient space for the registration task.

Let $\mathcal{F}$ be the space of real-valued, absolutely continuous functions on $[0,1]$ equipped with the so-called Fisher-Rao Riemannian metric. A priori, the metric is difficult to calculate. The Square-Root Velocity Function (SRVF) $q$ of $f \in \mathcal{F}$ is defined as $q : [0,1] \to \mathbb{R}$ where $\forall t \in [0,1]$,

$$q(t) \equiv \begin{cases} \frac{\dot{f}(t)}{\sqrt{|\dot{f}(t)|}} & \text{if } \left|\dot{f}(t)\right| \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Up to a translation, it is a one-to-one map. As $f$ is absolutely continuous, the resulting SRVF is square integrable ($\mathbb{L}^2$ is thus defined as the space of all SRVFs). Remarkably, it can be demonstrated that under the SRVF representation, the Fisher-Rao metric becomes the standard $\mathbb{L}^2$ metric. The next step is to consider an equivalence class of $q \in \mathbb{L}^2$, denoted $[q]$. Any two elements of $[q]$ represent functions which have the same amplitude variability but different phase variability. The quotient space of $\mathbb{L}^2$ under this equivalence relation is denoted $\mathcal{S} = \mathbb{L}^2/\Gamma$ where $\Gamma$ is the set of orientation-preserving diffeomorphisms of the unit interval $[0,1]$. An elastic distance is defined on $\mathcal{S}$. For $f_1, f_2 \in \mathcal{F}$, their corresponding SRVFs $q_1, q_2 \in \mathbb{L}^2$, the elastic distance $d$ is defined as

$$d([q_1],[q_2]) = \inf_{\gamma \in \Gamma} \left\| q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}} \right\|. \tag{3}$$

Finding the optimal registration for $f_1$ and $f_2$ is actually the same as computing the elastic distance (see Srivastava and Klassen (2016), Definition 4.7, p.99). From Equation 3, it is clear that elastic registration is not simply a least-square alignment of SRVFs. The Karcher mean (also known as the Fréchet mean) on $\mathcal{S}$ is used to derive a template for the registration. All details are provided by Anuj Srivastava et al. 2011.

To execute elastic alignment, it entails selecting a component $d$ of the trajectory characterized by distinct inflections signifying the transition between flight phases. Unlike the longitude profile, the altitude profile happens to exhibit the appropriate characteristics as the succession of flight phases is delineated by distinct breaks, as illustrated on Figure 2. The altitude profile obtained through elastic alignment is presented in Figure 3. Time warping functions are obtained using the Dynamic Programming (DP) algorithm presented by Srivastava and Klassen (2016) (Appendix B, p.435) and implemented by Tucker (2024). The average altitude profile now reveals the plateaus of the approach phase. Interestingly, as shown in Figure 4, once elastic registration is performed, landmarks almost perfectly coincide with the template chosen in the landmark registration procedure.

# 4 Conclusion and perspectives

The statistical analysis of aircraft trajectories requires a pre-processing step. This necessity arises from the very fact that, in general, flights are neither simultaneous nor of the same duration. They exhibit strong operational variations. We must ensure to compare comparable

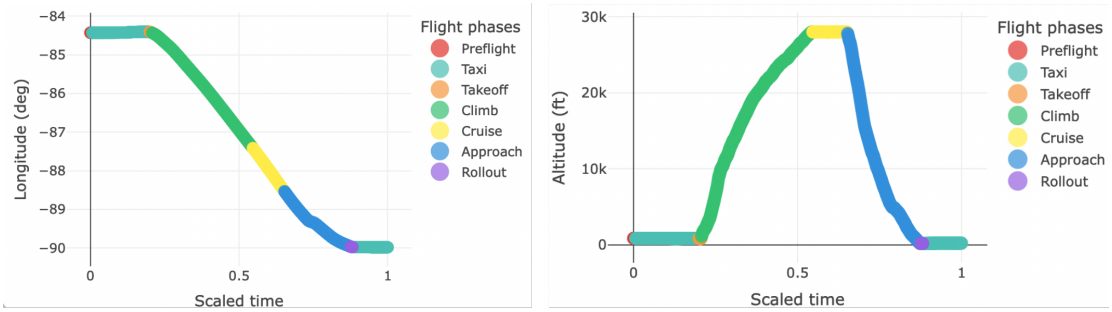



Figure 2: For trajectory n°2, it is evident that the longitude values do not exhibit any inflection points associated with a transition from one flight phase to another. Any elastic alignment procedure relying on longitude values would not yield an average altitude amplitude profile as desired, meaning for similar flight phases. We will then use the altitude profile.

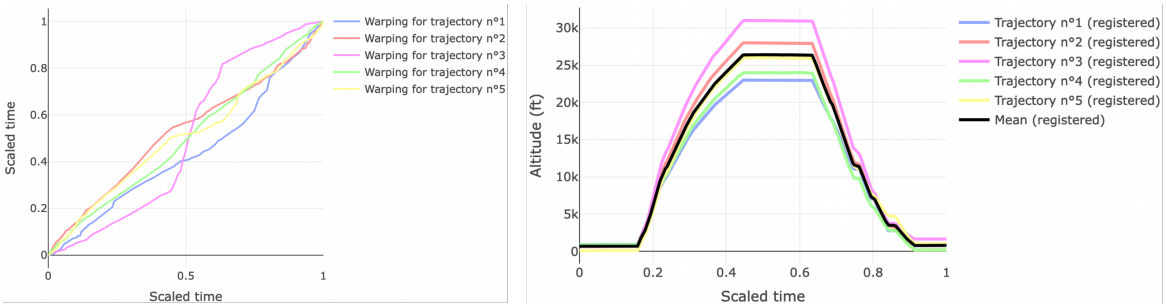



Figure 3: Time warping functions [left] and aligned trajectories [right] when using elastic registration.

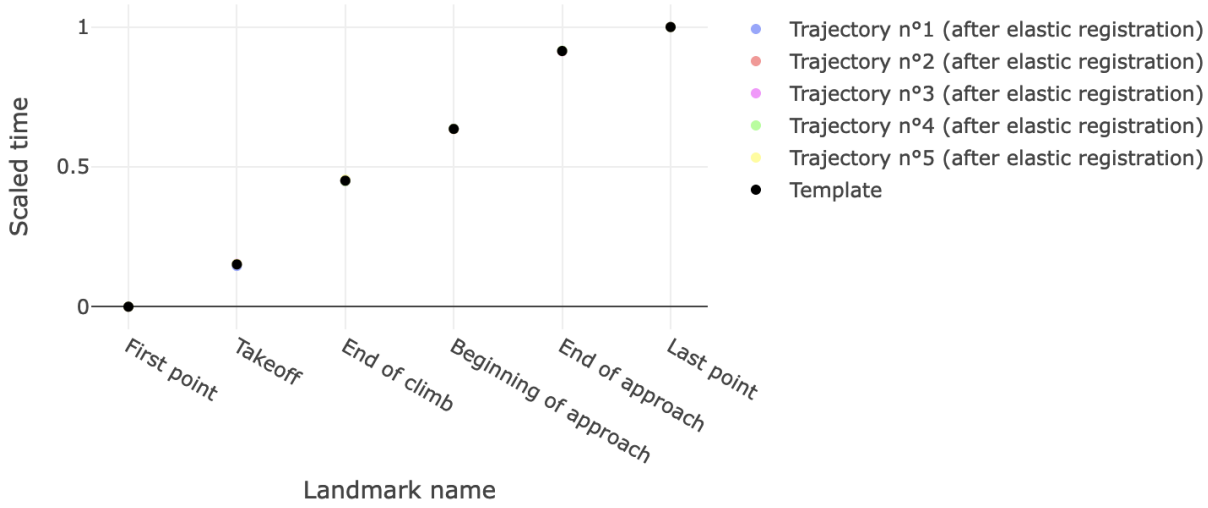



Figure 4: Once elastic alignment is performed, the landmarks almost perfectly coincide with the template chosen in the landmark alignment procedure.

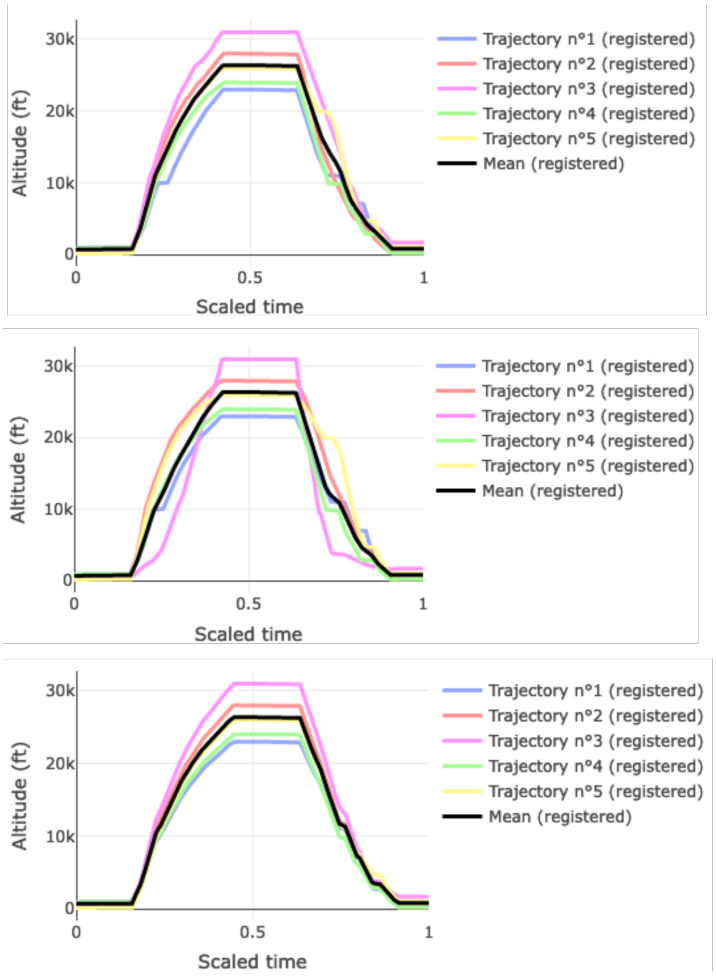



Figure 5: Landmark registration when warping functions are based on linear interpolation [top], monotone cubic Hermite spline interpolation [middle] and elastic registration [bottom].

flight instances, meaning, ensuring that the flight phases coincide after the registration. The simplest approach, when flight phases are known, is to perform landmark-based registration. However, even in this ideal scenario, it remains that landmark registration is only discrete evidence concerning intrinsically continuous warping functions. In the example presented, elastic registration, more complex both conceptually and computationally, allows for obtaining a more detailed average altitude profile, as summarized on Figure 5. Its reliability relies on selecting a component where we have good reason to believe that the breakpoints reflect the transition from one flight phase to another.

This last point is not so obvious for general aviation or drone flights. Several components are usually necessary to characterize flight phases, which are more numerous and more difficult to segment. A drone may, for instance, hover in place while rotating. A natural extension of elastic registration in this case lies in the alignment of parametric curves for which the theoretical framework has been proposed by Srivastava et al. 2011.

# References


Andrieu, Cindie, Baptiste Gregorutti, Florence Nicol, and Stéphane Puechmorel. 2016. "Espaces de courbes pour l'analyse de données aéronautiques." 48ème Journées de Statistique, Montpellier.

Delahaye, D., S. Puechmorel, P. Tsiotras, and E. Feron. 2014. "Mathematical Models for Aircraft Trajectory Design: A Survey." In *Air Traffic Management and Systems,* 205–247. Lecture Notes in Electrical Engineering. Tokyo: Springer Japan. ISBN: 978-4-431-54475-3. https://doi.org/10.1007/978-4-431-54475-3_12.

Fala, Nicoletta, Georgios Georgalis, and Nastaran Arzamani. 2023. "Study on Machine Learning Methods for General Aviation Flight Phase Identification." *Journal of Aerospace Information Systems* 20 (10): 636–647. ISSN: 1940-3151. https://doi.org/10.2514/1.I011246.

Gasser, Theo, and Alois Kneip. 1995. "Searching for Structure in Curve Sample." *Journal of the American Statistical Association* 90, no. 432 (December): 1179. ISSN: 01621459. https://doi.org/10.2307/2291510.

Kneip, Alois, and Theo Gasser. 1992. "Statistical Tools to Analyze Data Representing a Sample of Curves." *The Annals of Statistics* 20, no. 3 (September). ISSN: 0090-5364. https://doi.org/10.1214/aos/1176348769.

Kokoszka, Piotr, and Matthew Reimherr. 2021. *Introduction to functional data analysis.* Texts in statistical science series. Boca Raton London New York: CRC Press. ISBN: 978-1-03-209659-9 978-1-4987-4634-2.

Koner, Salil, and Ana-Maria Staicu. 2023. "Second-Generation Functional Data." *Annual Review of Statistics and Its Application* 10 (1): 547–572. https://doi.org/10.1146/annurev-statistics-032921-033726.

Marron, J. S., James O. Ramsay, Laura M. Sangalli, and Anuj Srivastava. 2014. "Statistics of time warpings and phase variations." *Electronic Journal of Statistics* 8, no. 2 (January): 1697–1702. ISSN: 1935-7524, 1935-7524. https://doi.org/10.1214/14-EJS901.

———. 2015. "Functional Data Analysis of Amplitude and Phase Variation." *Statistical Science* 30, no. 4 (November). ISSN: 0883-4237. https://doi.org/10.1214/15-STS524.

Marron, J.S., and Ian L. Dryden. 2021. *Object Oriented Data Analysis.* Boca Raton: Chapman / Hall/CRC, October. ISBN: 978-1-351-18967-5.

Müller, Hans-Georg. 2016. "Peter Hall, functional data analysis and random objects." *The Annals of Statistics* 44, no. 5 (October). ISSN: 0090-5364. https://doi.org/10.1214/16-AOS1492.

Nicol, Florence. 2013. "Functional principal component analysis of aircraft trajectories." In *2nd International Conference on Interdisciplinary Science for Innovative Air Traffic Management (ISIATM).*

Perrichon, Rémi, Xavier Gendre, and Thierry Klein. 2024. "Hidden Markov Models and Flight Phase Identification." *Journal of Open Aviation Science* 1, no. 1 (January). ISSN: 2773-1626. https://doi.org/10.59490/joas.2024.7269.

Puechmorel, S., and D. Delahaye. 2007. "4D Trajectories: A functional data perspective." In *2007 IEEE/AIAA 26th Digital Avionics Systems Conference*. October. https://doi.org/10.1109/DASC.2007.4391832.

Ramsay, J. O. 1982. "When the data are functions." *Psychometrika* 47, no. 4 (December): 379–396. ISSN: 1860-0980. https://doi.org/10.1007/BF02293704.

Ramsay, J. O., and C. J. Dalzell. 1991. "Some Tools for Functional Data Analysis." *Journal of the Royal Statistical Society. Series B (Methodological)* 53 (3): 539–572. ISSN: 0035-9246.

Ramsay, J. O., and B. W. Silverman. 2005. *Functional Data Analysis*. Springer Series in Statistics. Springer New York. ISBN: 978-0-387-40080-8 978-0-387-22751-1.

Srivastava, A, E Klassen, S H Joshi, and I H Jermyn. 2011. "Shape Analysis of Elastic Curves in Euclidean Spaces." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, no. 7 (July): 1415–1428. ISSN: 0162-8828. https://doi.org/10.1109/TPAMI.2010.184.

Srivastava, Anuj, and Eric P. Klassen. 2016. *Functional and Shape Data Analysis*. Springer Series in Statistics. Springer New York. ISBN: 978-1-4939-4018-9 978-1-4939-4020-2.

Srivastava, Anuj, Wei Wu, Sebastian Kurtek, Eric Klassen, and J. S. Marron. 2011. *Registration of Functional Data Using Fisher-Rao Metric,* May.

Suyundykov, Ruslan, Stéphane Puechmorel, and Louis Ferré. 2010. "Multivariate Functional Data Clusterization by PCA in Sobolev Space Using Wavelets." 42ème Journées de Statistique, Marseille.

Tucker, J. Derek. 2024. *CRAN package 'fdasrvf' (version 2.2.0)*. https://github.com/jdtuck/fdasrvf_R.

Wang, Haonan, and J. S. Marron. 2007. "Object oriented data analysis: Sets of trees." *The Annals of Statistics* 35, no. 5 (October): 1849–1873. ISSN: 0090-5364, 2168-8966. https://doi.org/10.1214/009053607000000217.

Wang, Jane-Ling, Jeng-Min Chiou, and Hans-Georg Müller. 2016. "Functional Data Analysis." *Annual Review of Statistics and Its Application* 3 (1): 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624.

# Données manquantes

# HSMM piloté par les observations pour l'estimation de la dynamique des adventices

Hanna Bacave[1] & Nikolaos Limnios[2] & Nathalie Peyrard[1]

[1] *INRAE, UR MIAT, Université de Toulouse, Castanet-Tolosan, France.* {*hanna.bacave, nathalie.peyrard*}*@inrae.fr*

[2] *Sorbonne University Alliance, Unversité de Technologie de Compiègne, LMAC, France. nikolaos.limnios@utc.fr*

**Résumé.** Les adventices sont des plantes qui poussent spontanément dans les parcelles agricoles et qui entrent en compétition avec les cultures. Leur dynamique repose sur la colonisation et la dormance. La banque de graines n'étant jamais observée de manière naturelle, une modélisation de cette dynamique a été proposée dans le cadre des Hidden Markov Models (HMM). Ce modèle, appelé Observation Driven-HMM (OD-HMM) étend les HMM au cas où les probabilités de transition dépendent de l'observation courante pour tenir compte des nouvelles graines produites qui entrent dans la banque de graines. Cependant, pour plus de réalisme sur la distribution de la survie de la banque de graines, le cadre naturel serait celui des Hidden Semi-Markov Models (HSMM). Néanmoins la notion de durée de séjour dans l'état caché n'est plus adaptée dès lors que l'observation influence la chaîne cachée à chaque instant. En nous appuyant sur les deux cadres OD-HMM et HSMM, nous proposons un nouveau modèle général : l'OD-HSMM, permettant à la fois de tenir compte d'une influence des données sur la chaîne cachée et de s'affranchir de la loi du temps de séjour géométrique. Nous en présentons une version paramétrique à partir des paramètres clés de la dynamique d'une espèce adventice et nous discutons différentes approches pour leur estimation.

**Mots-clés.** HSMM, OD-HMM, algorithme ABC, algorithme EM, adventices

**Abstract.** Weeds are plants that grow spontaneously in agricultural plots and compete with crops. Their dynamics are based on colonization and dormancy. As the seed bank is never observed, a model of these dynamics has been proposed within the HMM framework. This model, called OD-HMM, extends HMM to the case where transition probabilities depend on current observation, to take account of newly produced seeds entering the seed bank. However, for more realism on the survival distribution of the seed bank, the natural framework would be Hidden Semi-Markov Models (HSMM). However, the notion of sojourn time in the hidden state is no longer appropriate, since observation influences the hidden chain at every instant. Combining the OD-HMM and HSMM frameworks, we propose a new general model, the OD-HSMM, which allows us to take into account the influence of data on the hidden chain, while at the same time going beyond the geometric distribution of sojourn time. We present a parametric version based on key parameters of the dynamics of a weed species, and discuss different approaches for their estimation.

**Keywords.** HSMM, OD-HMM, ABC algorithm, EM algorithm, weeds

# 1 Introduction

Les adventices sont des plantes qui poussent spontanément dans les parcelles agricoles et qui entrent en compétition avec les cultures, entraînant parfois une baisse du rendement. Dans le même temps, elles constituent un refuge pour les insectes pollinisateurs, comme les abeilles, ce qui leur confèrent un rôle important dans l'équilibre écologique. Ainsi, il est important de comprendre et donc de modéliser leur développement pour trouver un mode de gestion offrant un compromis entre conservation et éradication des plantes. La difficulté vient du fait qu'un élément important de la dynamique des plantes en général et des adventices en particulier est invisible : la banque de graines dans le sol. En général on ne dispose de données que sur la flore levée. Dans ce contexte avec données manquantes, ou cachées, une modélisation de type modèle de Markov caché (Hidden Markov Model, HMM) a naturellement été proposée (Pluntz et al., 2018). Ce modèle étend les HMM au cas où les probabilités de transition dépendent de l'observation courante pour tenir compte des nouvelles graines produites qui entrent dans la banque de graines. Pour cette raison il est appelé Observation-Driven HMM (OD-HMM, Bacave et al., 2023).

Cependant, dans l'OD-HMM la chaîne cachée vérifie toujours l'hypothèse markovienne et son utilisation implique donc de supposer que la durée de vie de la banque de graines est distribuée selon une loi géométrique, réduisant considérablement la probabilité d'avoir une durée de vie de la graine importante. Or, dans la réalité les graines peuvent survivre plusieurs dizaines d'années dans le sol (Baskin and Baskin, 2014). Le cadre naturel pour une loi du temps de séjour plus générale est le cadre des Hidden Semi Markov Models (HSMM, Barbu and Limnios, 2008; Yu, 2016). En nous appuyant sur ce cadre, nous proposons donc un nouveau modèle général, l'OD-HSMM, permettant à la fois de tenir compte d'une influence des données sur la chaîne cachée et de sortir de la loi du temps de séjour géométrique (Section 2).

Nous présentons ensuite une version paramétrique de l'OD-HSMM permettant d'intégrer les paramètres en jeu dans la dynamique d'une espèce adventice, à savoir ses probabilités de colonisation, germination, grenaison ainsi que de survie (Section 3).

Enfin, nous explorons deux approches pour l'estimation des paramètres. La première repose sur l'algorithme EM (Dempster et al., 1977), classiquement utilisé pour les modèles à variables cachées. Nous l'appliquons ici à l'OD-HSMM non paramétrique car l'étape M est alors explicite pour la mise à jour des probabilités de transition et d'émission puis nous rajoutons une étape de minimisation pour identifier les paramètres adventices à partir de ces distributions. La seconde repose sur l'algorithme Approximate Bayesian Computation (ABC, Sisson et al., 2019) permettant d'estimer directement la distribution des paramètres. Nous discutons les avantages et limites de chacune de ces deux méthodes dans le cadre de l'estimation des paramètres en jeu dans la dynamique de l'adventice (Section 4).

# 2 Extension du cadre HSMM au cas où la chaîne cachée est pilotée par les observations

## 2.1 Rappel sur les HSMM

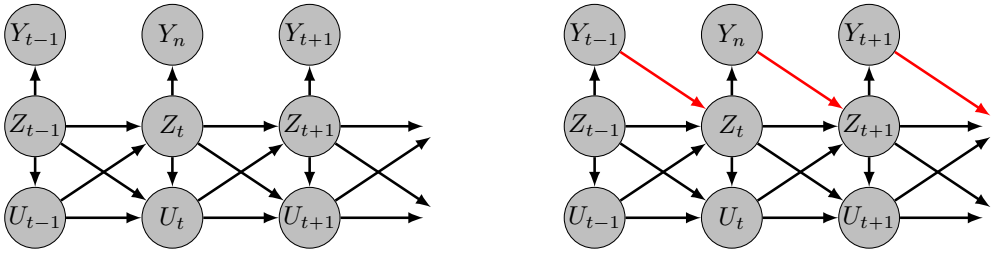
Nous considérons ici le cas temps discret et espaces d'états discrets. Un modèle de semi-Markov caché (Hidden Semi-Markov Model, HSMM) se compose de deux processus aléatoires, indexés par le temps : le processus observé $Y = (Y_t)_{t\in\mathbb{N}}$, dont l'espace d'états est $\Omega_Y = \{1, ..., D\}$, et le processus caché $Z = (Z_t)_{t\in\mathbb{N}}$, qui est une chaîne de semi-Markov avec pour espace d'états $\Omega_Z = \{1, ..., S\}$. Les observations entre $t = 0$ et $t = M$ sont désignées par le vecteur $Y_{0:M} = (Y_0, Y_1, Y_2, ..., Y_M)$. De même, le vecteur des états cachés entre $t = 0$ et $t = M$ est noté $Z_{0:M} = (Z_0, Z_1, ..., Z_M)$. Ainsi, dans la version classique du HSMM, la distribution jointe de $(Z_{0:M}, Y_{0:M})$ est entièrement déterminée par les distributions suivantes : la probabilité initiale $\mathbb{P}(Z_0 = i)$ (notée $\pi(i)$), la probabilité d'émission $\mathbb{P}(Y_t = a|Z_t = i)$ (notée $R(i, a)$) et enfin, le noyau de transition $\mathbb{P}(Z_{t+d} = j, Z_{t+1:t+d-1} = i|Z_t = i, Z_{t-1} \neq i)$ (noté $q_{ij}(d)$), où $d$ est la durée de séjour de la chaîne cachée dans l'état $i$. Le noyau de transition du HSMM est le produit de la probabilité que la durée de séjour de la chaîne cachée dans un état $i$ soit $d$, sachant que l'état suivant sera $j$, et de la probabilité de transition de cet état vers l'état $j$.

## 2.2 Comment introduire une dépendance aux observations dans un HSMM ?

Dans ce travail, on cherche à étendre le HSMM classique au cas où l'observation $Y_{t-1}$ a une influence sur la variable cachée suivante $Z_t$. Dans ce cas, on ne peut plus définir une variable aléatoire représentant la durée de séjour conditionnellement à l'entrée dans l'état puisqu'à chaque instant l'impact de l'observation peut venir modifier cette durée. La définition d'un HSMM à partir du noyau $q_{ij}(d)$ ne semble donc pas adaptée. Il existe une autre expression d'un HSMM, basée sur l'introduction d' un nouveau processus $U = (U_t)_{t\in\mathbb{N}}$ décrivant le temps passé depuis l'entrée dans l'état $Z_t$, avec comme espace d'états $\Omega_U = \mathbb{N}^*$. Ainsi, alors que $Z_t$ est une chaîne de semi-Markov, le couple $(Z_t, U_t)$ forme une chaîne de Markov (Barbu and Limnios, 2008). Cette modélisation d'un HSMM est donc plus adaptée pour une extension au cas où les observations $(Y_t)$ ont une influence, à chaque pas de temps, sur les états cachés $(Z_t, U_t)$. C'est celle que nous adoptons.

## 2.3 Définition générale du modèle OD-HSMM

Le modèle OD-HSMM (Observation-Driven HSMM) est composé de trois processus aléatoires : le processus observé $Y = (Y_t)_{t\in\mathbb{N}}$, le processus caché $Z = (Z_t)_{t\in\mathbb{N}}$ et le processus $U = (U_t)_{t\in\mathbb{N}}$ (caché également) des temps passés depuis l'entrée dans le dernier état. Soit $U_{0:M} = (U_0, U_1, U_2, ..., U_M)$. Dans un OD-HSMM la distribution jointe de $(Z_{0:M}, U_{0:M}, Y_{0:M})$ est définie à partir des distributions suivantes :




(a) HSMM (b) OD-HSMM

FIGURE 1 – Représentation graphique des dépendances conditionnelles dans la chaîne $(Z_t, U_t, Y_t)$ : (a) cadre HSMM, (b) cadre OD-HSMM.

— La loi initiale $\mathbb{P}(Z_0 = z_0, U_0 = 1) = \mathbb{P}(Z_0 = z_0)$ est notée $\pi(z_0)$ (on suppose que l'on entre dans un nouvel état à $t = 0$) ;
— La probabilité d'émission (qui reste indépendante de $U_t$ comme dans un HSMM) $\mathbb{P}(Y_t = y_t | Z_t = z_t, U_t = u_t) = \mathbb{P}(Y_t = y_t | Z_t = z_t)$ est notée $R(z_t, y_t)$ ;
— La probabilité de transition $\mathbb{P}(Z_t = z_t, U_t = u_t | Z_{t-1} = z_{t-1}, U_{t-1} = u_{t-1}, Y_{t-1} = y_{t-1})$ est notée $P_{y_{t-1}}((z_{t-1}, u_{t-1}), (z_t, u_t))$.

Les dépendances conditionnelles dans un OD-HSMM sont représentées dans la Figure 1.

**Remarque 1.** *En raison de l'influence des observations sur les états cachés, la probabilité de transition dépend de l'observation $y_{t-1}$. Donc pour $z_t, u_t, z_{t-1}, u_{t-1}$ fixés il y a autant de probabilité de transition à calculer que d'états observés possibles.*

**Remarque 2.** *La probabilité de transition est à définir quelque soit $z_t, u_t, z_{t-1}, u_{t-1}$. Cependant, seul un faible nombre de probabilités sont non nulles car $u_t$ ne peut valoir que $u_{t-1} + 1$ ou 1.*

# 3 Modélisation de la dynamique de l'adventice avec un OD-HSMM paramétrique

## 3.1 Paramètres en jeu dans la dynamique de l'adventice

L'adventice est une plante annuelle, c'est-à-dire qu'elle meurt chaque année. Les paramètres en jeu dans la dynamique d'une population d'adventices sont la germination, la grenaison, la colonisation et la survie de la banque de graines. En s'inspirant de la modélisation de type OD-HMM avec espaces d'états en présence/absence proposée par Pluntz et al. (2018), nous définissons les paramètres suivants à partir desquels nous allons construire les probabilités d'émission et de transition du OD-HSMM.

1. Une partie des graines dans la banque de graines peut germer et produire des plantes, avec probabilité $g$.
2. Indépendamment d'une possible germination, le stock de graines peut survivre dans le sol, mais cette survie diminue avec le temps passé dans le sol $u$. On définit donc la

probabilité que la banque de graines survive entre $t$ et $t+1$ si elle existe depuis déjà $u$ pas de temps par

$$s = s_0 e^{-\lambda(u-1)},$$

où $s_0$ est la probabilité qu'un nouveau stock de graines créé à $t$ survive à $t+1$ et $\lambda$ modélise la vitesse avec laquelle la banque de graine s'épuise.

3. Une fois que les graines ont poussé et sont devenues des plantes en fleur, ces dernières dispersent leurs graines au sol de la parcelle. Celles-ci entrent donc dans la banque de graines. C'est ce qu'on appelle la grenaison, qui est de probabilité $d$.
4. Chaque année, la quantité de graines dans le sol peut augmenter grâce à la colonisation par des graines provenant d'un autre champ. On suppose une colonisation de probabilité constante, $c$.

On notera $p_0$ la probabilité qu'il y ait déjà des graines dans le sol la première année d'observation.

## 3.2 Lois du modèle

Comme dans Pluntz et al. (2018), nous nous plaçons dans le cadre présence/absence (i.e. avec $\Omega_Z = \Omega_Y = \{0,1\}$). Soit $\theta = (p_0, c, s_0, d, g, \lambda)$ le vecteur des paramètres, les lois de l'OD-HSMM s'écrivent de la façon suivante en fonction de $\theta$.

La loi initiale $\pi(z_0)$ s'exprime en fonction de $p_0$ : $(\pi(0), \pi(1)) = (1-p_0, p_0)$.

La loi d'émission $R(z_t, y_t)$ représente la germination des graines d'adventices. Elle s'exprime donc en fonction de $g$, de la façon suivante :

$$R = \begin{pmatrix} 1 & 0 \\ 1-g & g \end{pmatrix},$$

où le premier élément de la première ligne désigne la probabilité $\mathbb{P}(Y_t = 0 | Z_t = 0)$, qui vaut 1 puisqu'il ne peut pas y avoir de plante en fleur lorsqu'il n'y a pas de graine dans le sol.

La probabilité de transition dépend de $u_{t-1}$ et $u_t$ qui ne sont pas bornés. Pour faciliter la présentation, nous présentons les différentes valeurs prises par $P_{y_{t-1}}((z_{t-1}, u_{t-1}), (z_t, u_t))$ via une « matrice » de transition décomposée en 4 blocs selon les valeurs de $z_{t-1}$ et $z_t$ même si ces blocs sont de taille infinie ($|\Omega_U| \times |\Omega_U|$). Ces 4 blocs correspondent au 4 états possibles du couple $(Z_t, U_t)$, ils sont notés $A^k_{i,j}$, où $\forall u_{t-1}, u_t,\ A^k_{i,j}(u_{t-1}, u_t) = P_k(i, u_{t-1}, j, u_t)$. On exprime les deux « matrices » de transition en fonction des $A^k_{i,j}$ comme suit :

$$P_0 = \left(\begin{array}{c|c} A^0_{0,0} & A^0_{0,1} \\ \hline A^0_{1,0} & A^0_{0,1} \end{array}\right)$$

$$P_1 = \left(\begin{array}{c|c} A^1_{0,0} & A^1_{0,1} \\ \hline A^1_{1,0} & A^1_{0,1} \end{array}\right)$$

Avec :

$$A_{ij}^k = \begin{cases} \begin{pmatrix} 0 & \alpha_i^k(1) & 0 & \dots \\ 0 & 0 & \alpha_i^k(2) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} & \text{si } i = j, \\ \begin{pmatrix} \beta_i^k(1) & 0 & 0 & \dots \\ \beta_i^k(2) & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix} & \text{si } i \neq j. \end{cases}$$

Où $\forall k \in \{0,1\}$ :

$$\alpha_i^0(u) = \begin{cases} (1-c) \text{ si } i = 0 \\ 1-(1-c)(1-s_0 \times e^{-\lambda(u-1)}) \text{ si } i = 1 \end{cases},$$

$$\alpha_i^1(u) = \begin{cases} (1-c)(1-d) \text{ si } i = 0 \\ 1-(1-c)(1-s_0 \times e^{-\lambda(u-1)})(1-d) \text{ si } i = 1 \end{cases},$$

et $\forall k \in \{0,1\},\ \beta_i^k(u) = 1 - \alpha_i^k(u)$.

**Example 1** (Détail du calcul de $\alpha_1^0(2)$)**.**

$$\begin{aligned} \alpha_1^0(2) = & \quad \mathbb{P}(Z_t = 1, U_t = 3 | Z_{t-1} = 1, U_{t-1} = 2, Y_{t-1} = 0) \\ = & \quad 1 - \mathbb{P}(Z_t = 0, U_t = 1 | Z_{t-1} = 1, U_{t-1} = 2, Y_{t-1} = 0), \end{aligned}$$

*où la probabilité* $\mathbb{P}(Z_t = 0, U_t = 1 | Z_{t-1} = 1, U_{t-1} = 2, Y_{t-1} = 0)$ *décrit l'événement dans lequel il n'y a plus de graine dans le sol alors qu'il y en avait précédemment depuis deux pas de temps mais qu'il n'y avait pas de flore levée. Ainsi, les graines dans le sol n'ont pas survécu, événement dont la probabilité vaut* $(1 - s_0 \times e^{-\lambda(2-1)})$ *et il n'y a pas non plus eu de colonisation, ce qui est de probabilité* $(1 - c)$*.*

## 3.3 Exemple

Afin de rendre plus compréhensible le lien entre les paramètres et les probabilités de transition, nous présentons ici les 3 premières lignes et colonnes des blocs $A_{i,j}^k$. Pour alléger

les notations, on pose $s_u = s_0 \times e^{-\lambda(u-1)}$.

$$P_0 = \left(\begin{array}{cccc|cccc} 0 & 1-c & 0 & \dots & c & 0 & 0 & \dots \\ 0 & 0 & 1-c & \dots & c & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots & c & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \hline (1-c)(1-s_1) & 0 & 0 & \dots & 0 & 1-(1-c)(1-s_1) & 0 & \dots \\ (1-c)(1-s_2) & 0 & 0 & \dots & 0 & 0 & 1-(1-c)(1-s_2) & \dots \\ (1-c)(1-s_3) & 0 & 0 & \dots & 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{array}\right)$$

$$P_1 = \left(\begin{array}{cccc|cccc} 0 & (1-c)(1-d) & 0 & \dots & 1-(1-c)(1-d) & 0 & 0 & \dots \\ 0 & 0 & (1-c)(1-d) & \dots & 1-(1-c)(1-d) & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots & 1-(1-c)(1-d) & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \hline (1-c)(1-s_1)(1-d) & 0 & 0 & \dots & 0 & 1-(1-c)(1-s_1)(1-d) & 0 & \dots \\ (1-c)(1-s_2)(1-d) & 0 & 0 & \dots & 0 & 0 & 1-(1-c)(1-s_2)(1-d) & \dots \\ (1-c)(1-s_3)(1-d) & 0 & 0 & \dots & 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{array}\right)$$

# 4 Estimation des paramètres de l'OD-HSMM pour adventice

Nous explorons deux approches : l'approche EM, naturelle dans le cas d'un modèle avec données manquantes ainsi que l'approche ABC, adaptée au cas où la vraisemblance n'est pas accessible.

## 4.1 Approche EM

L'algorithme EM repose sur la quantité intermédiaire suivante, où $\theta^{(m)}$ est la valeur courante du paramètre :

$$Q(\theta|\theta^{(m)}) = \mathbb{E}\left[\ln \mathbb{P}(Y_{0:M}, Z_{0:M}, U_{0:M}|\theta)|Y_{0:M} = y_{0:M}, \theta^{(m)}\right]$$

Il s'agit d'un algorithme itératif où chaque itération est composée de deux étapes : à l'étape E on calcule les distributions marginales intervenant dans l'expression de la quantité intermédiaire, puis à l'étape M on met à jour l'ensemble des paramètres en résolvant $\theta^{(m+1)} = \arg\max_{\theta} Q(\theta|\theta^{(m)})$.

Dans le cas de l'OD-HSMM pour estimer la dynamique de l'adventice, il n'existe pas de formule explicite pour mettre à jour les paramètres $c$, $d$, $s_0$ et $\lambda$ lors de l'étape M. Pour palier à ça, il est possible de résoudre le problème de maximisation de façon numérique en utilisant, par exemple, des outils disponibles dans R. Quelques essais préliminaires n'ont pas conduit à des résultats satisfaisants, ainsi nous explorons une autre approche. Dans le cas non paramétrique, l'étape M de mise à jour des probabilités de transition et d'émission $(P_y, R)$

conduit à des expressions analytiques simples. Nous choisissons donc de mettre en oeuvre l'EM sur l'OD-HSMM non paramétrique et d'ajouter une étape supplémentaire d'identification des paramètres $\theta$ à partir des $P_y$ et $R$ estimés, à l'issue du EM.

### 4.1.1 Étape E

L'étape E de l'algorithme EM pour l'OD-HSMM est très similaire à celle de l'algorithme EM pour HMM (mais avec une variable cachée bi-dimensionnelle) et repose sur l'algorithme Forward-Backward. Cependant, la récursion Backward a été modifiée pour prendre en compte le fait que la probabilité de transition dépend de l'observation.
A priori l'expression de $Q(\theta|\theta^{(m)})$ faut intervenir des sommes infinies sur $U_t$ et $U_{t-1}$ mais en pratique ces variables sont bornées par $M$ puisque l'on suppose qu'à $t = 0$ la chaîne cachée entre dans un nouvel état.

### 4.1.2 Étape M

L'étape M consiste à mettre à jour l'ensemble des probabilités d'émission et de transition, noté $\theta^{NP} = (\{P_y((i,u),(j,u'))\}, \{R(i,a)\})$. Pour cela, on résout le problème de maximisation $\arg\max_{\theta^{NP}} Q(\theta^{NP}|\theta^{NP(m)})$. Les formules de mise à jour sont explicites. Par exemple pour $R(1,1)$ on obtient

$$R(1,1)^{(m+1)} = \frac{\sum_{t=0}^{M} \mathbb{P}(Z_t = 1|Y_{0:M} = y_{0:M})\mathbb{1}_{y_t=1}}{\sum_{t=0}^{M} \mathbb{P}(Z_t = 1|Y_{0:M} = y_{0:M})}.$$

### 4.1.3 Identification des paramètres

Le paramètre $g$ étant égal à $R(1,1)$ son estimation est directe à partir des estimateurs $\hat{\theta}^{NP} = (\{\hat{P}_y((i,u),(j,u'))\}, \{\hat{R}(i,a)\})$ obtenus à l'issue du EM. Soit $P_y(\theta)$ l'expression des probabilités de transition en fonction des paramètres $(c, s_0, d, \lambda)$ du modèle adventice (cf Section 3). L'étape d'identification des paramètres consiste à minimiser la distance entre $\hat{P}_y$ et $P_y(\theta)$, grâce à des outils classiques de résolution de systèmes non linéaires.

## 4.2 Approche ABC

L'algorithme ABC est un algorithme d'estimation bayésienne qui consiste à générer un grand nombre de valeurs pour les paramètres $\theta$ du modèle et à identifier ceux qui conduisent à des données simulées « proches » de celles observées. En pratique, pour une valeur échantillonnée $\hat{\theta}$, la première étape consiste à générer des données simulées selon le modèle. Puis, dans une seconde étape des statistiques sont calculées sur ces données simulées et comparées à celles calculées sur les observations réelles. Dans le cas où la distance entre

les deux est inférieure à un certain seuil, les paramètres $\hat{\theta}$ utilisés pour la simulation sont acceptés. L'ensemble des valeurs $\hat{\theta}$ retenues fournit une distribution a posteriori de $\theta$.

Pour mettre en oeuvre ABC pour estimer les paramètres de l'OD-HSMM adventice, nous utilisons le package R « EasyABC » (Jabot et al., 2013). Nous proposons d'utiliser les statistiques suivantes calculées sur $y_{0:M}$ :

— le nombre de 0 ;
— le minimum, le maximum, la moyenne et les quantiles à 25%, 50% et 75% des longueurs des phases de 1 consécutifs et des phases de 0 consécutifs de 0 et 1 ;
— le nombre de transitions de 0 vers 1.

Un avantage à utiliser l'algorithme ABC est que cette méthode ne nécessite pas de savoir calculer la vraisemblance, elle repose uniquement sur la simulation du modèle, qui est très simple pour le modèle OD-HSMM. Enfin, l'algorithme ABC est une méthode d'estimation bayésienne donc il permet d'obtenir une estimation de la distribution des paramètres plutôt qu'un estimateur ponctuel comme avec l'algorithme EM. En revanche, choisir les bonnes statistiques peut s'avérer être un exercice compliqué. C'est ce que nous sommes en train d'explorer sur des données simulées.

# 5 Conclusion

Nous proposons un modèle général, l'OD-HSMM, permettant de combiner les avantages du cadre HSMM et du cadre OD-HMM. Ce modèle est motivé par une application à la modélisation de la dynamique des adventices mais il peut avoir aussi un intérêt dans d'autres domaines comme la finance (Engel and Hamilton, 1990) pour prédire le régime d'un système monétaire en fonction du taux de change.

Nous décrivons ensuite une version paramétrique de l'OD-HSMM, pour la dynamique d'une adventice. Nous comparons deux méthodes d'estimation. La première consiste à estimer les lois du modèle général non paramétrique grâce à l'algorithme EM, puis à identifier les paramètres adventice par résolution d'un système non linéaire. La seconde estime la distribution des paramètres par le biais de l'algorithme ABC. A ce stade ABC nous semble plus adapté du fait de la simplicité de mise en oeuvre, mais des expérimentations sur données simulées sont en cours afin de comparer les performances des algorithmes EM et ABC. Ensuite, l'algorithme le plus performant sera appliqué à un jeu de données réel sur les adventices (Pluntz et al., 2018).

Une perspective à ce travail est d'étendre le modèle à un cadre spatial plus réaliste en modélisant explicitement la colonisation entre parcelles au lieu de supposer qu'il s'agit d'une pluie de graines. Pour cela, nous explorerons l'extension de l'OD-HSMM au cas multi-chaînes où chaque chaîne décrit la dynamique de l'adventice dans une parcelle donnée. La modélisation de l'influence de la flore levée d'une parcelle sur la banque de graines d'une autre pose les mêmes difficultés de modélisation que l'OD-HSMM, à savoir qu'une variable extérieure vient influer sur le temps de séjour de la chaîne locale. Une piste consistera donc à ré-exploiter la formulation d'un HSMM par le couple $(Z, U)$. L'extension de l'OD-HSMM

au cas multi-chaînes pourra être comparée à la modélisation de Le Coz et al. (2019) qui est moins générale car dans le cadre HMM et non HSMM.

# Références


Bacave, H., P.-O. Cheptou, N. Limnios, and N. Peyrard (2023). Non parametric Observation Driven HMM. Preprint available at https://hal.inrae.fr/hal-04053732v1.

Barbu, V.-S. and N. Limnios (2008). *Semi-Markov Chains and Hidden Semi-Markov Models towards Applications*. Springer.

Baskin, C. and J. Baskin (2014). *Seeds : Ecology, Biogeography, and, Evolution of Dormancy and Germination*. San diego, Academic Press.

Dempster, A., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*, 1–22.

Engel, C. and J.-D. Hamilton (1990). Long swings in the dollar : Are they in the data and do markets know it ? *The American Economic Review 80*(4), 689–713.

Jabot, F., T. Faure, and N. Dumoulin (2013). EasyABC : performing efficient approximate Bayesian computation sampling schemes using R. *Methods in Ecology and Evolution*, 684–687.

Le Coz, S., P. O. Cheptou, and N. Peyrard (2019). A spatial Markovian framework for estimating regional and local dynamics of annual plants with dormancy. *Theoretical Population Biology 127*, 120–132.

Pluntz, M., S. Le Coz, N. Peyrard, R. Pradel, R. Coquet, and P. O. Cheptou (2018). A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora. *Ecology Letters 21*, 1311–1318.

Sisson, A. S., F. Yanan, and M. A. Beaumont (2019). *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC.

Yu, S.-Z. (2016). *Hidden Semi-Markov Models Theory, Algorithms and Applications*. Elsevier.

# La distribution Ex-Gauss pour l'analyse du temps de réaction : Initialisation plus robuste et traitement des données manquantes

Alandra Zakkour [1,2], Yousri Slaoui [1] & Cyril Perret [1,2]

[1] *Laboratoire de Mathématiques et Applications, Université de Poitiers, France.*
[2] *Centre de Recherches sur la Cognition et l'Apprentissage, Université de Poitiers, France.*
*E-mail : alandra.zakkour@univ-poitiers.fr*
*E-mail : yousri.slaoui@univ-poitiers.fr*
*E-mail : cyril.perret@univ-poitiers.fr*

**Résumé.** Le temps entre la présentation d'un stimulus et la réponse motrice d'un participant est la mesure la plus ancienne et la plus largement utilisée pour explorer le fonctionnement de l'esprit humain. Donders a théorisé cette durée, appelée temps de réaction (RT), comme impliquant trois ensembles d'activités : les mécanismes perceptifs, le traitement cognitif et la préparation motrice. Partant de l'hypothèse que les premier et dernier ensembles de traitements peuvent être considérés comme ayant des durées quasi identiques pour une même tâche, tout changement de RT entre deux conditions expérimentales est alors interprété comme indiquant un changement de la durée des traitements cognitifs. RTs sont alors considérées par les psychologues comme un outil pour explorer les mécanismes de traitement cognitif.
Pour analyser cette mesure, nous nous référons à la distribution Ex-Gauss, largement étudiée dans la littérature. Notre étude propose une méthode permettant d'obtenir des estimations moins biaisées pour les trois paramètres de cette distribution ($\mu$, $\sigma$, et $\tau$) en utilisant une approche bayésienne. Cette méthode consiste à adapter l'initialisation des paramètres en recourant au rééchantillon-nage de Bootstrap plutôt qu'à une sélection aléatoire de vecteurs de paramètres initiaux.
Un deuxième aspect essentiel de ce travail est la résolution du problème des données manquantes de type MAR, caractérisées par différents pourcentages de présence.
**Mots-clés.** RT; Distribution Ex-Gauss; Initialisation des paramètres; Données Manquantes; MAR

**Abstract.** The time between the presentation of a stimulus and a participant's motor response is the oldest and most widely used measure for exploring the functioning of the human mind. Donders theorized this duration, called reaction time (RT), as involving three sets of activities: perceptual mechanisms, cognitive processing, and motor preparation. Assuming that the first and last sets of processes can be considered to have nearly identical durations for the same task, any change in RT between two experimental conditions is then interpreted as indicating a change in the duration of cognitive processing. RTs are thus considered by psychologists as a tool for exploring cognitive processing mechanisms.
To analyze this measure, we refer to the Ex-Gaussian distribution, widely studied in the literature. Our study proposes a method to obtain less biased estimates for the three parameters of this distribution ($\mu$, $\sigma$, and $\tau$) using a Bayesian approach. This method involves adapting

the parameter initialization by resorting to Bootstrap resampling instead of randomly selecting initial parameter vectors.
A second essential aspect of this work is the resolution of the missing data problem of the MAR type, characterized by different percentages of presence.
**Keywords.** RT; Ex-Gaussian Distribution; Parameter Initialization; Missing Data; MAR

# 1 Introduction

Plusieurs distributions de probabilité ont été proposées pour prendre en compte le temps de réaction (RT). La distribution de probabilité qui semble être la plus proche de la distribution observée est obtenue en convoluant une distribution gaussienne et une distribution exponentielle, c'est-à-dire une distribution Ex-Gaussienne ( Burbeck (1982) ; Hohle (1965) ; Luce (1986) ; El Haj et al., (2021)).
El Haj et al., (2021) ont proposé une méthode bayésienne pour estimer les trois paramètres Ex-Gaussiens, $\mu$, $\sigma$ et $\tau$. L'objectif de cette méthode était d'obtenir des estimations non biaisées dans les conditions restrictives d'un petit échantillon ($n < 100$), fréquemment observées dans les études de psychologie scientifique. Cette méthodologie bayésienne utilise une initialisation aléatoire des paramètres.

Le premier objectif de notre article est d'adopter cette méthode en utilisant une initialisation appropriée dans l'approche bayésienne basée sur le rééchantillonnage et plus spécifiquement sur le Bootstrap.
Le bootstrap est une approche qui peut fournir des estimations précises, particulièrement lorsque les approximations habituelles sont invalides. Cette technique implique la création de multiples échantillons bootstrap en sélectionnant aléatoirement des observations à partir de l'ensemble de données original avec remplacement. Chaque échantillon a la même taille que les données initiales, permettant aux observations d'être incluses plusieurs fois ou pas du tout. Pour plus de détails sur cette méthode et leurs applications pratiques, voir le tutoriel de Wehrens (2000) ; Davison (1997).

Le deuxième objectif de ce travail est de traiter le problème des données manquantes de type Missing At Random (MAR), qui est le cas le plus classique défini par Rubin (1976) où la valeur manquante est prédite uniquement sur la base des données observées.
La méthode d'Imputation Multiple (MI) est devenue l'une des approches les plus avancées pour aborder le problème des données manquantes. C'est une technique statistique basée sur la création de plusieurs ensembles de valeurs possibles pour les données manquantes. L'imputation multiple est un cas général de l'Imputation Simple (SI) dans lequel les données manquantes sont remplacées par une valeur possible, puis les paramètres sont estimés. Pour obtenir la méthode d'imputation multiple, nous répétons la méthode simple plusieurs fois avec différentes valeurs prédites, puis les résultats sont combinés. MI aide à produire des inférences statistiques plus précises et fiables par rapport à la SI.
Dans notre article, nous avons suggéré de comparer la méthode MI avec trois autres méthodologies,

qui sont documentées dans la littérature.

Pour valider nos propositions, nous avons illustré son application en utilisant à la fois des données simulées et réelles. La comparaison est réalisée en utilisant l'Augmentation du Risque Absolu "Absolute Risk Increase" (`ARI`) pour les vecteurs de paramètres, obtenus à travers notre méthode proposée et l'approche précédemment utilisée.

# 2 Methodologie

## 2.1 La distribution Ex-Gauss

Supposons que la variable aléatoire $Z$ suit la loi ex-gaussienne. Elle peut donc s'écrire sous la forme de la somme de deux variables aléatoires $X$ et $Y$:

$$Z = X + Y$$

où $X \sim \mathrm{N}(\mu, \sigma)$ et $Y \sim \exp(\lambda)$. Cette distribution possède la fonction de densité suivante:

$$f(x; \mu, \sigma, \lambda) = \frac{\lambda}{2} \exp\left(\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)\right)\phi\left(\frac{\mu + \lambda\sigma^2 - x}{\sqrt{2}\sigma}\right) \tag{1}$$

Avec $\phi$ est la fonction d'erreur complémentaire définit par:

$$\phi(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2)dt$$

Pour estimer les trois paramètres $\mu$, $\sigma$ et $\lambda$, El Haj et al., (2021) ont proposé une méthode basée sur l'inférence bayésienne pour produire des prédictions avec de petits échantillons. Notre étude présente une version plus efficace de cette méthode, dans laquelle le choix du vecteur initial est basé sur l'algorithme Bootstrap plutôt que d'être effectué de manière aléatoire.
Dans la section suivante, nous présentons la méthode de rééchantillonnage utilisée.

## 2.2 Bootstrap

La méthode de bootstrap est une technique de rééchantillonnage largement utilisée en statistiques pour estimer la distribution d'un échantillon statistique en se basant sur les données disponibles. Elle consiste à générer de multiples échantillons bootstrap en tirant aléatoirement avec remplacement des observations à partir de l'échantillon original. Ces échantillons bootstrap sont de taille égale à l'échantillon original (Efron and Tibshirani (1993)).

Soit $X = \{X_1, \ldots, X_n\}$ notre échantillon initial de taille n avec une fonction de distribution $F(x)$. Pour générer un échantillon bootstrap, nous effectuons un tirage aléatoire de n observations de l'échantillon initial avec remplacement. Cette procédure est répétée B fois (où

B est le nombre d'itérations). L'échantillon bootstrap obtenu est noté $X^* = \{X_1^*, \ldots, X_n^*\}$.

Pour chaque échantillon bootstrap, nous calculons la fonction de distribution $F^*(x)$ et l'estimateur de l'intérêt $\hat{\theta}^* = \{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$.

En résumé, la méthode Bootstrap permet d'approximer la distribution de l'estimateur statistique $\hat{\theta}$ en utilisant des échantillons bootstrap et de fournir des estimations robustes.

L'objectif est de sélectionner un vecteur de paramètres initiaux plus robuste qu'un vecteur arbitraire. Notre vecteur comprend trois paramètres à estimer, notés $\hat{\theta} = \{\hat{\mu}, \hat{\sigma}, \hat{\lambda}\}$. La méthode de rééchantillonnage est alors appliquée à trois fonctions d'intérêt statistique (voir l'algorithm 1).

Pour les applications numériques, nous avons comparé cette méthode proposée avec la

**Algorithm 1** L'algorithm Bootstrap: $X$ est le vecteur initial; $B$ est le nombre d'échantillons Bootstrap; $\theta = \{\mu, \sigma^2, \lambda\}$ est le vecteur de paramètre.

**Input:** $X$, $B$, $\mu$, $\sigma^2$ and $\lambda$.

1: **for** $b = 1, \ldots, B$ **do**

2: $X_b^* =$ échantillon de $X$ avec remplacement et de taille n;
$\mu_b^* = \mathbb{E}(X_b^*) - \left(0.5 * \sqrt{\mathbb{V}(X_b^*)}\right)$;
$(\sigma^2)_b^* = \mathbb{V}(X_b^*) - \left(0.5 * \sqrt{\mathbb{V}(X_b^*)}\right)^2$;
$\lambda_b^* = \frac{1}{0.5 * \sqrt{\mathbb{V}(X_b^*)}}$.

3: **end for**

**output:** $\mu^* = \frac{1}{B}\sum_{b=1}^{B} \mu_b^*$, $(\sigma^2)^* = \frac{1}{B}\sum_{b=1}^{B} (\sigma^2)_b^*$ and $\lambda^* = \frac{1}{B}\sum_{b=1}^{B} \lambda_b^*$.

méthode arbitraire de El Haj et al, (2021) ainsi qu'avec la méthode du maximum de vraisemblance sur trois exemples de données simulées de taille n=50. Les résultats obtenus démontrent l'efficacité de l'approche Bootstrap en se basant sur le critère d'Augmentation absolue du risque ($ARI$). Dans ce document nous avons présenté un seul exemple dans le tableau 1.

Dans la figure 1, nous présentons la convergence des paramètres de la distribution ex-gaussienne vers les valeurs initiales lorsque la taille de l'échantillon augmente.

| | Data | Aléa. | Max.V. | B |
|---|---|---|---|---|
| $\mu$ | **400** | 72.2533 | 497.6375 | **397.1385** |
| $\sigma$ | **200** | 42.77476 | **203.7802** | 181.9681 |
| $\frac{1}{\lambda}$ | **100** | 420.4193 | 12.17656 | **80.19481** |
| $ARI$ | **0** | 4.8096 | 1.1412 | **0.2953** |

Table 1: Prédiction des paramètres pour un jeu de données simulées de valeur $\mu = 400$, $\sigma = 200$, et $\frac{1}{\lambda} = 100$ en utilisant l'inférence bayésienne avec 5000 itérations. "Aléa" désigne la méthode d'initialisation aléatoire; "Max.V." désigne la méthode du maximum de vraisemblance; et "B" désigne la méthode de rééchantillonnage "Bootstrap".

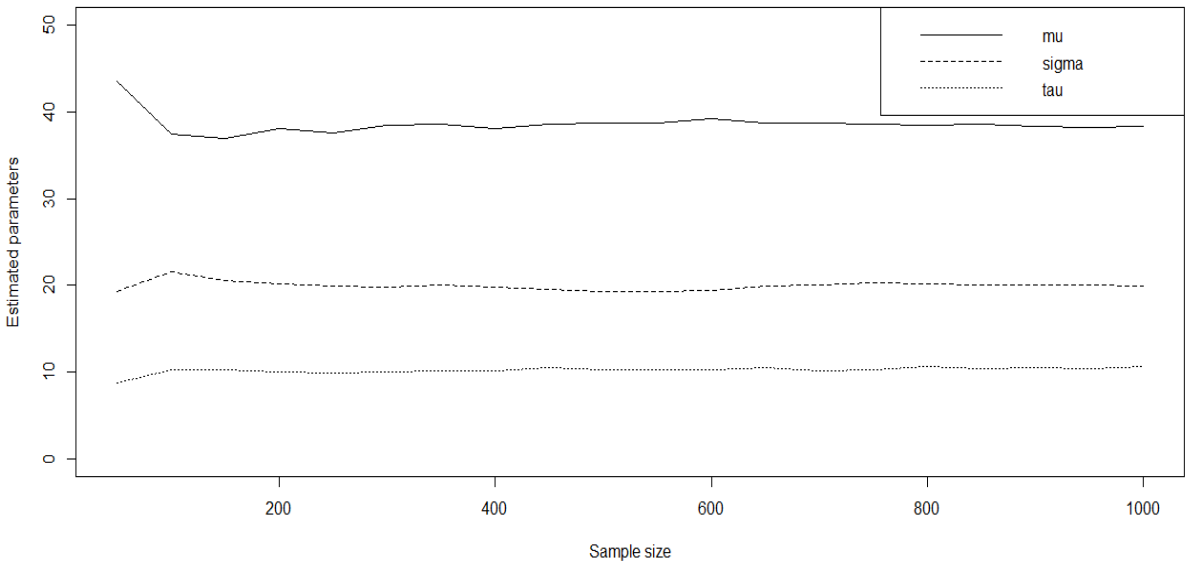



Figure 1: Exemple des paramètres estimés en utilisant la méthode Bootstrap pour différentes tailles d'échantillons à partir d'un jeu de données avec des valeurs initiales de $\mu = 40$, $\sigma = 20$ et $\frac{1}{\lambda} = 10$.

Notre deuxième objectif dans cette étude concerne le problème des données manquantes de type MAR, où la présence des données manquantes est associée à certaines caractéristiques observées. Dans la section suivante, nous abordons cette problématique ainsi que les méthodes que nous proposons pour y remédier.

## 2.3 Missing Data Imputation

Nous pouvons parler de données incomplètes lorsque les valeurs dans notre vecteur de réponse ne sont pas toutes observées pour de nombreuses raisons; on peut également dire qu'il s'agit d'une question sans réponse. Ces types de données sont un problème courant reconnu par les statisticiens.
En général, les données manquantes se produisent lorsqu'une valeur dans les données n'est pas représentée pour une variable donnée, pour de nombreuses raisons qui peuvent être liées à l'objectif de l'étude (par exemple, les participants ne répondent pas aux questions). Cela apparaît dans de nombreuses études de recherche, en particulier lors de la collecte de données et lorsque les participants sont étudiés sur une période de temps. Au début, les études étaient développées en supposant l'absence de valeurs manquantes. À la fin des années 1980, avec l'avancée de la technologie, ce problème a attiré l'attention de nombreux chercheurs qui souhaitaient étudier plusieurs techniques pour le gérer. En fonction des raisons de leur absence, ces valeurs pourraient être divisées en trois types : manquant complètement au hasard

| | | $\pi = 15\%$ | | | | $\pi = 20\%$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Data | Stand | K-NN | Bay | IM | Stand | K-NN | Bay | IM |
| $\mu$ | **4** | 3.691 | 3.613 | 3.531 | **3.739** | 3.738 | 3.678 | 3.589 | **3.818** |
| $\sigma$ | **2** | 1.447 | 1.436 | **1.627** | 1.548 | 1.440 | 1.469 | **1.627** | 1.584 |
| $\tau$ | **1** | 0.979 | **0.997** | 0.966 | 0.952 | **0.970** | 0.959 | 0.955 | 0.935 |
| *ARI* | **0** | 0.373 | 0.380 | **0.336** | 0.338 | 0.374 | 0.386 | 0.333 | **0.317** |

Table 2: Prédiction des paramètres pour un jeu de données incomplètes dans deux cas ($\pi$ c'est le pourcentage de présence des données manquantes).

(MCAR), manquant de manière aléatoire (MAR) et manquant de manière non aléatoire (MNAR).

Pour comprendre chaque type et leur différence voir les références Mack et al. (2018); Heitjan and Basu (1996). Dans cette étude, nous nous intéressons aux données manquantes de type MAR (la probabilité qu'une valeur soit manquantes dépend des données observées) et avons exploré plusieurs scénarios. Nous avons considéré quatre exemples de pourcentages de données incomplètes (5%, 10%, 15% et 20%) dans trois ensembles de données simulées. Les résultats ont été comparés entre quatre méthodes différentes. La méthode standard ("Stand") consiste à remplacer les données manquantes par la moyenne des données observées. La méthode des k plus proches voisins ("K-NN") prédit les valeurs manquantes en se basant sur les valeurs des voisins les plus proches. La méthode bayésienne ("Bay") remplace les données incomplètes en utilisant le théorème de Bayes. Enfin, la méthode d'imputation multiple ("IM") remplace les données non observées par des valeurs numériques obtenues par imputation simple, cette opération étant répétée plusieurs fois pour obtenir une imputation multiple.

Après avoir fixé la méthode de Bootstrap pour le choix du vecteur initial dans la méthode bayésienne pour l'estimation des paramètres ($\hat{\mu}$, $\hat{\sigma}$, $\hat{\lambda}$), nous avons appliqué les quatre méthodes proposées pour résoudre le problème des données manquantes (Na) pour les quatre pourcentages et sur les trois exemples de données simulées. Ensuite, le même phénomène a été répété sur quatre exemples de données réelles en psycholinguistique. Les résultats obtenus montrent une compétition entre les méthodes, notamment entre la méthode bayésienne et l'imputation multiple. Cela est observé dans le tableau 2, où nous avons présenté un exemple de jeu de données ($\mu = 4$, $\sigma = 2$ et $\lambda = 1$) avec deux pourcentages de valeurs manquantes (15% et 20%). Pour un taux de 15% de données manquantes, la méthode bayésienne présente la plus petite valeur du critère ARI (0.336), mais cette valeur est très proche de celle obtenue pour l'imputation multiple (0.338). Tandis que pour un taux de 20% de valeurs manquantes, le meilleur résultat est obtenue avec la méthode "IM", avec une valeur de 0.317, tandis que la méthode "Bay" présente une valeur de 0.333.

En répétant l'exemple sur plusieurs scénarios, avec différents taux de valeurs manquantes et plusieurs jeux de données, les résultats obtenus nous permettent d'envisager une autre méthode plus robuste consistant à imputer les données manquantes dans le cadre de l'estimation bayésienne.

# 3 Conclusion

Notre projet aborde deux problématiques distinctes. Dans un premier temps, notre objectif initial était de développer une méthode d'initialisation des paramètres permettant d'obtenir des estimations plus robustes et moins biaisées dans le cadre bayésien.
Les résultats obtenus montrent que la méthode de rééchantillonnage proposée, nommée Bootstrap, s'avère plus efficace que la méthode d'initialisation arbitraire utilisée par El Haj et al,. (2021).
Ensuite, nous avons exploré la résolution du problème des données manquantes de type MAR en variant les pourcentages de présence des données. Cette partie de notre étude n'a pas permis d'identifier de manière concluante une méthode prédominante, car les méthodes bayésiennes et d'imputation multiple ont produit des résultats similaires. Face à cette compétition, nous ouvrons une nouvelle voie de recherche pour trouver une méthode permettant de résoudre ce problème de manière plus efficace.

# References


Burbeck, S. and Luce, R. (1982), Evidence from auditory simple reaction times for both change and level detectors, *Perception & psychophysics*, 32, pp. 117–133.

Davison, A. and Hinkley, D. (1997). Bootstrap Methods and Their Applications. *Cambridge Univ. Press, Cambridge.*

Donders, F. C. (1869/1969), On the speed of mental processes, *Acta Psychologica*, 30, pp. 412-431.

Efron, B. and Tibshirani, R. (1993), An Introduction to the Bootstrap, *Chapman & Hall/CRC, New York.*

El Haj, A., Slaoui, Y., Solier, C., and Perret, C. (2021), Bayesian Estimation of The Ex-Gaussian Distribution, *Stat. Optim. Inf. Comput. 9*, pp 809-819.

Enders, C.K. (2010), Applied Missing Data Analysis, *Guilford Press: New York, NY, USA.*

Heitjan, D. and Basu, Srabashi. (1996), Distinguishing "Missing at Random" and "Missing Completely at Random". *Amer. Statist.*, 50, pp 207-213.

Hohle, R.H. (1965), Inferred components of reaction times as function of foreperiod duration, *Journal of Experimental Psychology*, 69, pp 382-386.

Little, R.J. and Rubin, D.B. Statistical Analysis with Missing Data, (2019), *John Wiley & Sons: Hoboken, NJ, USA*, 793.

Luce, R. D. (1986), Response times: Their role in inferring elementary mental organization, *Oxford: Oxford University Press.*

Mack, C., Su, Z. and Westreich, D. (2018), Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user's guide. *Agency for Healthcare Research and Quality (US).*

Schafer, L.J. and Graham, W.J. (2002), Missing Data: Our View of the State of the Art, *Psychol. Methods*, 7, pp. 147–177.

Roelofs, A. (2018), One hundred fifty years after Donders : Insignts form unpublished data, a replication, and modeling of his reaction times, *Acta Psychologica*, 191, pp. 228-233.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika.*, 63, pp. 581-590.

Wehrens, R., Putter, H. and Buydens, L. (2000). The bootstrap: a tutorial. *Chemometr. Intell. Lab. Syst.*, 54, pp. 35-52.

# Subspace clustering sur données incomplètes

Yasmine Agliz [1,2] & Vincent Audigier [3] & Ndèye Niang [4]

[1] *Laboratoire CEDRIC, équipe MSDMA, CNAM, Fance, yasmine.agliz@caissedesdepots.fr*
[2] *Caisse des dépôts, DAJCD, France*
[3] *Laboratoire CEDRIC, équipe MSDMA, CNAM, Fance, vincent.audigier@cnam.fr*
[4] *Laboratoire CEDRIC, équipe MSDMA, CNAM, Fance, ndeye.niang_keita@cnam.fr*

**Résumé.** Nous proposons ici une nouvelle approche pour la classification non-supervisée sur données incomplètes dans le cadre d'un grand nombre de variables. Cette approche intitulée *Reduced K-pod* s'appuie sur la formulation d'un critère de Reduced K-means calculable sur des données incomplètes sur le modèle de la méthode *K-pod*. Un algorithme d'optimisation du critère est proposé et sa convergence monotone est garantie. Cette méthode est ensuite évaluée par une étude par simulation mettant en évidence l'apport de la méthode par rapport aux approches géométriques concurrentes gérant soit les données manquantes (*K-pod*), soit les données manquantes et la grande dimension (par ACP itérative suivie de K-means). Les premiers résultats obtenus indiquent de meilleures performances en termes d'indice de Rand Ajusté mettant ainsi en évidence l'intérêt de l'approche pour la gestion de la grande dimension et des données manquantes en classification.

**Mots-clés.** Subspace clustering, classification, données grande dimension, données manquantes au hasard.

**Abstract.** A new approach for clustering on incomplete data within the framework of a large number of variables is proposed. This approach, entitled *Reduced K-pod*, is based on the formulation of a Reduced K-means criterion computable on incomplete data, in lines with the *K-pod* method. An algorithm for optimising the criterion is proposed with monotonic convergence ensured. This method is then evaluated using a simulation study highlighting the properties of the method compared with competitive geometric clustering approaches handling either missing data (*K-pod*) or missing data and high dimensionality (by iterative PCA followed by K-means). The preliminary obtained results indicate better performances in terms of Adjusted Rand Index, highlighting the interest of the approach for addressing high dimensionality and missing data in clustering.

**Keywords.** Subspace clustering, clustering, high dimensional data, Missing At Random data.

# 1 Introduction

Nous nous intéressons au problème de la classification automatique d'individus décrits par un grand ensemble de variables, ceci en nous focalisant sur approches géométriques. Dans ce cadre de grande dimension, de nombreuses variables deviennent non pertinentes pour la tâche de classification et les classes sont alors souvent décrites par des sous-espaces de variables. Or, ces variables non-pertinentes peuvent pénaliser l'apprentissage des algorithmes classiques de classification [1, 2].

Pour pallier à cette inefficacité, les algorithmes de *Subspace Clustering* ont notamment été proposés. L'objectif de ces derniers est de retrouver des classes et leur sous-espaces caractéristiques. Ces derniers peuvent être obtenus à travers des méthodes de sélection de variables basées sur des systèmes de pondération [3] ou par combinaisons linéaires de l'ensemble des variables. Parmi ces dernières, on peut citer l'*approche tandem* [4], consistant à appliquer la classification sur les premières composantes principales d'une analyse factorielle. Toutefois, cette approche reste critiquable car si les composantes principales maximise l'inertie de projection du nuage de points, rien ne garantit qu'elle maximise la dispersion des centres de gravité des classes recherchées [5, 6]. Une autre approche consiste alors à rechercher simultanément la partition des individus et les composantes optimales pour la tâche de classification comme proposé dans les méthodes Reduced K-means [5] et Factorial K-means [6] (en abrégé, *RKM* et *FKM* respectivement).

Par ailleurs, le grand nombre de variables rend incontournable le problème des données manquantes, constituant ainsi une difficulté supplémentaire. La tâche de classification par les approches géométriques dans le cadre de données incomplètes a cependant fait l'objet de différents travaux, notamment en "petite dimension". Une première approche, généralement peu recommandée, consiste à se ramener à un jeu de données complet en supprimant les observations ou variables incomplètes, ou en remplaçant les valeurs manquantes par imputation simple. D'autres approches, plus performantes, basées sur l'imputation multiple [7] ont également été proposées. On retrouve aussi dans la littérature des approches consistant à adapter les critères des méthodes de classification au caractère incomplet des données, comme proposé dans le cadre du k-means via l'approche *K-pod* [8], ou dans le cadre du fuzzy c-means et du k-prototype [9, 10].

Dans ce travail, nous nous intéresserons précisément à la classification sur données incomplètes dans le cadre d'un grand nombre de variables. Pour cela, nous définissons d'une part un nouveau critère, similaire à celui optimisé dans la méthode Reduced K-means, mais calculable sur données incomplètes, et d'autre part, un algorithme d'optimisation convergent de façon monotone, inspiré de celui de l'approche *K-pod*. On se place dans le cadre classique où les données sont manquantes au hasard [11] aussi appelées données *(Missing At Random)* (MAR).

La deuxième section présente les méthodes RKM et FKM, tandis que la troisième section présente la méthode *K-pod*. La méthode proposée, dénommée Reduced *K-pod*, fait l'objet de la quatrième section. Enfin, celle-ci est évaluée en cinquième section via une étude par simulation.

## 2 *Subspace clustering* par RKM et FKM

Ces deux approches peuvent être vues comme des adaptations de la populaire méthode des K-means dans un contexte de *subspace clustering*. La méthode des K-means peut en effet se présenter comme le problème de minimisation du critère suivant :

$$KM(\mathbf{C}, \mathbf{U}|c) = \min_{\mathbf{C},\mathbf{U}} ||\mathbf{X} - \mathbf{UC}||_F^2 \tag{1}$$

où $\mathbf{X}$ est la matrice des données de dimensions $(I \times J)$ comportant en ligne les individus et en colonne les variables, $\mathbf{U}$ la matrice d'appartenance des observations aux $c$ classes, composée de 0 et 1, de dimensions $(I \times c)$, $\mathbf{C}$ la matrice de centroïdes dans l'espace initial de dimensions $(c \times J)$, et $||.||_F$ la norme de Frobenius.

La méthode RKM consiste alors à reformuler ce critère de façon à ce que la matrice $\mathbf{C}$ s'exprime sous la forme d'un produit matriciel $\mathbf{FA}^\top$ où $\mathbf{F}$ de dimensions $(c \times q)$ est la matrice des centroïdes dans un espace réduit de dimension $q$ et $\mathbf{A}$ de dimensions $(J \times q)$ est une matrice de loadings déterminant la contribution de chaque variable à la structure en groupe des observations. On obtient alors le critère suivant :

$$RKM(\mathbf{A}, \mathbf{F}, \mathbf{U}|c, q) = \min_{\mathbf{U},\mathbf{A}} ||\mathbf{X} - \mathbf{UFA}^T||_F^2 \tag{2}$$

D'un point de vue modélisation, la méthode peut aussi se présenter selon

$$\mathbf{X} = \mathbf{UFA}^T + \mathbf{E} \tag{3}$$

avec $\mathbf{E}$ une matrice $(I \times J)$ de résidus indépendants identiquement distribués selon une loi normale centrée.

Là où le Reduced K-means vise à minimiser la somme des distances au carré entre les observations et les centroïdes de l'espace réduit, reconstitués dans l'espace initial, le Factoriel K-means lui consiste à minimiser la somme des distances au carré entre les observations dans l'espace réduit et leurs centroïdes correspondants dans ce même espace. Le critère de la méthode s'énonce alors ainsi :

$$FKM(\mathbf{A}, \mathbf{F}, \mathbf{U}|c, q) = \min_{\mathbf{U},\mathbf{A},\mathbf{F}} ||\mathbf{XA} - \mathbf{UF}||_F^2 \tag{4}$$

L'optimisation des critères (2) et (4) s'effectuent en alternant entre la recherche de la partition, via la matrice $\mathbf{U}$, obtenue par K-means, et la mise à jour du sous-espace, via les matrices $\mathbf{F}$ et $\mathbf{A}$, obtenues par décomposition en valeurs singulières. L'algorithme s'arrête lorsque le critère ne décroît plus. Afin d'éviter la convergence vers un minimum local, les algorithmes nécessitent plusieurs initialisations.

Les deux approches ont des propriétés assez similaires, on pourra se référer à [12] pour une étude comparative. Notons que [2] propose une combinaison de ces deux approches, appelée *Generalized Reduced Clustering*, via la définition d'un critère d'optimisation exprimé comme combinaison linéaire des critères (3) et (4).

# 3 K-means sur données incomplètes

Comme évoqué en introduction, la gestion des données manquantes en classification peut s'effectuer efficacement par les méthodes d'imputation multiple [11, 13] et les méthodes dites *directes*, qui consistent à adapter les méthodes de classification de façon à ce qu'elles s'accommodent des données manquantes.

Dans le cas de la méthode des K-means, la gestion des données manquantes par imputation multiple se déroule en trois étapes distinctes [7]. Dans un premier temps, le jeu de données est imputé M fois selon un modèle d'imputation respectant la structure en groupes des observations [14]. La deuxième étape consiste à appliquer l'algorithme des K-means sur chacun des tableaux imputés, fournissant ainsi un ensemble de $M$ partitions. Enfin, la troisième étape consiste en l'agrégation de ces dernières.

Concernant les méthodes "directes", [8] ont proposé de reformuler le critère (1) optimisé dans la méthodes K-means, de façon à ce qu'il puisse être évalué sur des données incomplètes. Celui-ci est défini comme suit :

$$f(\mathbf{U}, \mathbf{C}) = \min_{\mathbf{U},\mathbf{C}} \|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{UC})\|_F^2 \tag{5}$$

avec $\Omega \subset \{1, \ldots, I\} \times \{1, \ldots, J\}$ un sous-ensemble des indices correspondant aux entrées observées. L'opérateur de projection des matrices $I \times J$ sur l'ensemble d'indices $\Omega$ est défini comme suit :

$$[P_\Omega(X)]_{ij} = \begin{cases} x_{ij} & \text{si } (i,j) \in \Omega \\ 0 & \text{si } (i,j) \in \Omega^c \end{cases}$$

Ainsi, le critère (5) revient à rechercher la matrice d'appartenance $\mathbf{U}$ et les centroïdes associés, de façon à ce que les profils observés (et donc partiels) des individus soient les plus proches de ceux des centres associés. Parce qu'il ne porte donc pas sur les éléments manquants du profil, il est parfaitement calculable sur un jeu de données incomplet. La solution du critère peut être approchée par un algorithme de Majorization-Minimization (MM). Son principe consiste à alterner entre le K-means et l'imputation des observations incomplètes par les coordonnées de leur centroïde associé.

# 4 Reduced *K-pod*

Afin de proposer une nouvelle approche de classification en grande dimension sur données incomplètes, nous proposons un nouveau critère d'optimisation reprenant les critères précédents, spécifiques aux méthodes de subspace clustering et de classification sur données incomplètes. Plus précisément, celui-ci est défini selon les notations précédentes de la façon suivante :

$$f(\mathbf{U}, \mathbf{A}, \mathbf{F}) = \min_{\mathbf{U},\mathbf{A}} \|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{U}\mathbf{F}\mathbf{A}^T)\|_F^2 \tag{6}$$

Ce critère peut être exprimé comme une généralisation du critère (2) au cas incomplet. Son optimisation peut alors être effectuée à l'aide d'un algorithme MM où la fonction majorante $g$ au point courant $\tilde{U}, \tilde{F}, \tilde{A}$ se définit selon :

$$g(\mathbf{U}, \mathbf{F}, \mathbf{A}|\tilde{\mathbf{U}}, \tilde{\mathbf{F}}, \tilde{\mathbf{A}}) \;\; = f(\mathbf{U}, \mathbf{A}, \mathbf{F}) + \left\| P_{\Omega^c}(\mathbf{U}\mathbf{F}\mathbf{A}^{\mathbf{T}}) - P_{\Omega^c}(\tilde{\mathbf{U}}\tilde{\mathbf{F}}\tilde{\mathbf{A}}^{\mathbf{T}}) \right\|_F^2$$

et se minimise selon un algorithme RKM, assurant ainsi la convergence monotone de l'algorithme Reduced k pod (1).

**Algorithm 1** Reduced *K-pod*

**Entrées :** $\mathbf{X}$ un tableau de données incomplet, $c$ un nombre de classes, $q$ la dimension du sous-espace

**Initialiser** $\mathbf{F}^{(0)}$ et $\mathbf{A}^{(0)}$ par ACP itérative, puis mettre définir un tableau imputé $\mathbf{X}^{(0)}$ selon $\mathbf{X}^{(0)} \leftarrow P_\Omega(\mathbf{X}) + P_{\Omega^c}(\mathbf{F}^{(0)}\mathbf{A}^{(0)\top})$

**Pour** $\ell$ **de 1 à** $L-1$

- $(\mathbf{U}^{(\ell+1)}, \mathbf{F}^{(\ell+1)}, \mathbf{A}^{(\ell+1)}) \leftarrow$ Reduced k-means$(\mathbf{X}^{(\ell)})$
- $\mathbf{X}^{(\ell+1)} \leftarrow P_\Omega(\mathbf{X}) + P_{\Omega^c}(\mathbf{U}^{(\ell+1)}\mathbf{F}^{(\ell+1)}\mathbf{A}^{(\ell+1)\top})$

**Sorties :** $\mathbf{U}^{(L)}$, $\mathbf{F}^{(L)}$, $\mathbf{A}^{(L)}$

Cet algorithme consiste tout d'abord en une ACP itérative [15] afin d'obtenir les $q$ composantes principales $\mathbf{F}^{(0)}$ et vecteurs propres associés $\mathbf{A}^{(0)}$. Ceux-ci sont alors utilisés pour initialiser les valeurs manquantes de $\mathbf{X}$. Cette initialisation permet de tenir compte de la structure en $q$ dimension des données. Par la suite, l'algorithme alterne un RKM, afin de minimiser la fonction de coût $g$, et à mettre à jour les données manquantes de façon à pouvoir définir la valeur de la nouvelle fonction majorante au point courant. Ces étapes sont répétées jusqu'à atteindre un nombre prédéfini d'itérations $L$, ou arrêtées dès lors que le critère optimisé (6) ne décroît plus.

# 5 Simulations

## 5.1 Plan de simulation

Afin d'évaluer les performances de la méthode présentée, tant en termes d'identification de la partition que du sous-espace, nous nous plaçons dans le cadre d'un modèle de RKM (équation 3). Ceci implique de définir les trois matrices $\mathbf{U}$, $\mathbf{F}$, $\mathbf{A}$ et $\mathbf{E}_{\text{sim}}$. La matrice $\mathbf{X}$ est constituée de la sorte :

$$\mathbf{X} = \mathbf{UFA}^T + \mathbf{E}_{\text{sim}} \tag{7}$$

Pour cela, nous nous sommes inspirés du plan de simulation de [16]. Nous considérons une matrice $\mathbf{X}$ de dimensions $(I \times J)$, avec $p_1$ le nombre des variables informatives, $p_2$ le nombre des variables de bruit corrélées entres elles, et $p_3$ le nombre des variables de bruit indépendantes ($J = p1 + p2 + p3$). Dans ce plan, la matrice $\mathbf{U}$ est générée grâce à une loi multinomiale avec des probabilités égales. La matrice de centroïdes $\mathbf{F}$ est générée à partir d'une distribution uniforme $q$-dimensionnelle sur $[-15, 15]^q$. $\mathbf{A}$ est ensuite construite de la sorte :

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^{*T} & 0^T_{q\times(p2+p3)} \end{bmatrix}$$

avec $\mathbf{A}^*$ une matrice orthogonale de dimension $p1 \times q$ générée de manière aléatoire. Chaque élément de $\mathbf{E}_{\text{sim}}$, est généré à partir de la distribution normale $J$-dimensionnelle $\mathcal{N}(0, \Sigma_J)$, avec $\Sigma_J$ :

$$\Sigma_J = \begin{bmatrix} I_{p1} & 0_{p1\times p2} & 0_{p1\times p3} \\ 0_{p2\times p1} & \Sigma_{p2} & 0_{p2\times p3} \\ 0_{p3\times p1} & 0_{p3\times p2} & I_{p3} \end{bmatrix}$$

avec $\Sigma_{p2}$ la sous-matrice de dimensions $(p2 \times p2)$ de terme $\sigma_{ij}$ $(1 \leq i, j \leq p_2)$ valant 1 pour $i = j$ et 0.25 sinon.

Pour chacun des scénario, 200 jeux de données sont générés, en se basant sur des matrices $\mathbf{U}$, $\mathbf{F}$ et $\mathbf{A}$ fixes et en faisant varier $\mathbf{E}_{\text{sim}}$.
Pour chaque jeu de données, des données manquantes sont générées selon un mécanisme MCAR et MAR pour différents taux de données manquantes (5%, 15% et 25%). Sur les jeux de données complets, deux approches sont utilisées : RKM et l'approche tandem combinant l'ACP et le K-means sur les $q$ premières composantes principales. Pour RKM, les paramètres utilisés sont $c = 8$, $q = 2$. Pour l'*approche tandem*, les mêmes paramètres sont appliqués. Ensuite, pour les jeux de données incomplets, trois approches sont envisagées : Reduced *K-pod*, l'équivalent de l'*approche tandem* adaptée aux données incomplètes et le *K-pod*. L'*approche tandem* adaptée aux données incomplètes commence par une ACP itérative suivie d'une étape de K-means. Les mêmes paramètres que pour les données complètes sont utilisés pour la *tandem approche* et le Reduced *K-pod* : $c = 8$, $q = 2$. Pour le *K-pod*, le paramètre $c$ est fixé à 8.

Les performances des algorithmes sont mesurées selon l'indice de Rand ajusté (ARI) [4] entre les différentes partitions obtenues et la véritable partition des jeux de données. Pour Reduced K Means et Reduced *K-pod*, nous utiliserons le coefficient de congruence pour

comparer les matrices de *loadings* obtenues par les algorithmes et les matrices de *loadings* utilisées pour générer les données.

## 5.2 Résultats

Les Figures 1 et 2, représentent respectivement les ARI moyens sur les 200 jeux de données de ces deux scénarii ou $c = 8$, $n = 400$, $q = 2$ et $p_1 = p_2 = p_3 = 5$ ou 10.
Ici, nous constatons, dans un premier temps, une différence entre les ARI moyens entre les deux scénarii (avec un ARI moyen à 0.35 (Figure 1) et un ARI moyen 0.9 (Figure 2)) . Cela pourrait s'explique par une meilleure séparabilité des classes pour le deuxième scénario, observé sur le premier plan factoriel. Dans un second temps, on observe naturellement que plus le taux de données manquantes augmente, plus les performances des méthodes s'accommodant des données incomplètes diminuent. Pour la Figure 2, on constate que bien que les performances du Reduced *K-pod* décroissent plus rapidement, l'ARI du Reduced *K-pod* reste supérieure à celle de la tandem approche adaptée aux données manquantes. Dans les deux cas, le *K-pod* ne performe pas aussi bien que les deux autres méthodes. Cela peut s'expliquer par la structure en groupe qui se trouve dans un espace réduit ainsi que par l'initialisation des données manquantes par la moyenne des variables utilisée par la méthode du *K-pod*.

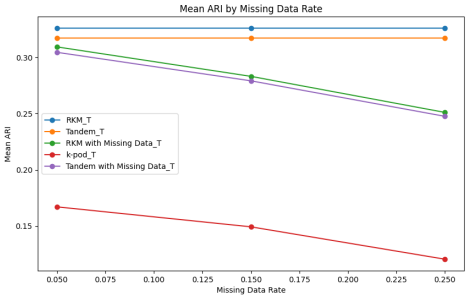



Figure 1: ARI moyen sur 200 jeux de données par taux de données manquantes (MCAR, $p_1 = p_2 = p_3 = 5$)

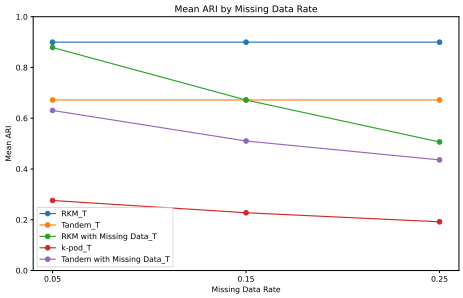



Figure 2: ARI moyen sur 200 jeux de données par taux de données manquantes (MCAR, $p_1 = p_2 = p_3 = 10$)

En ce qui concerne la Figure 3, on remarque que pour les différents taux de données manquantes, RKM performe en moyenne mieux que la *tandem approche* adaptée aux données manquantes pour chacun des jeux de données.

D'abord, on observe que la méthode Reduced *K-pod* a des valeurs d'ARI globalement supérieures à celles obtenues par une *approche tandem* avec ACP itérative.

En comparant les ARI obtenus selon Reduced *K-pod* et *K-pod*, tout comme précédemment, on observe que ces derniers sont plus élevés pour l'approche proposée. Cette différence plus marquée, peut s'expliquer par la prise compte de la structure en sous-espace. Cela peut

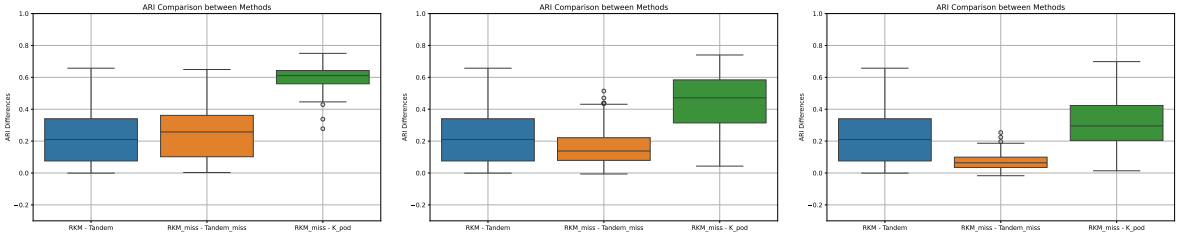



Figure 3: Distribution des différences d'ARI entre les différentes méthodes pour 5% (à gauche), 15% (au milieu) et 25% de données manquantes (à droite) selon un mécanisme MCAR, ($p_1 = p_2 = p_3 = 10$)

également s'expliquer par l'initialisation des données manquantes. En effet, la méthode du *K-pod* utilise une initialisation des valeurs manquantes par la moyenne des variables, ce qui peut altérer la structure en groupes des données.

# 6 Conclusion

Nous présentons ici une nouvelle méthode de classification sur données incomplètes dans le cadre d'un grand nombre de variables. Cette méthode s'appuie sur la formulation d'un critère tenant compte ces deux caractéristiques. L'algorithme d'optimisation associé converge de façon monotone. Cette première étude par simulation permet de mettre en évidence les meilleures performances de l'approche Reduced *K-pod* par rapport aux méthodes concurrentes (*K-pod* et *approche tandem*). Comme attendu, on constate une détérioration des performances avec le taux de données manquantes.

Cependant, ces performances sont à relativiser dans la mesure où le nombre de variables considéré dans cette simulation reste modeste. Une étude complémentaire est actuellement menée dans ce sens. D'autres travaux méthodologiques sont également en cours pour la prise en compte des données mixtes, à travers l'adaptation de la méthode des K-prototypes à la classification en grande dimension [10].

# References

[1] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. SIGKDD Explor., 6:90–105, 2004.

[2] Michio Yamamoto and Heungsun Hwang. A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika, 41(1):115–129, 2014.

[3] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. Journal of the American Statistical Association, 105(490):713–726, 2010.

[4] Lawrence Hubert and Phipps Arabie. Comparing partitions. Journal of classification, 2:193–218, 1985.

[5] Geert De Soete and J Douglas Carroll. K-means clustering in a low-dimensional euclidean space. In New approaches in classification and data analysis, pages 212–219. Springer, 1994.

[6] Maurizio Vichi and Henk AL Kiers. Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis, 37(1):49–64, 2001.

[7] Vincent Audigier and Ndèye Niang. Clustering with missing data: which equivalent for rubin's rules? Advances in Data Analysis and Classification, pages 1–35, 2022.

[8] Jocelyn T Chi, Eric C Chi, and Richard G Baraniuk. k-pod: A method for k-means clustering of missing data. The American Statistician, 70(1):91–99, 2016.

[9] R. J. Hathaway and J. C. Bezdek. Fuzzy c-means clustering of incomplete data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31(5):735–744, 2001.

[10] Rabea Aschenbruck, Gero Szepannek, and Adalbert FX Wilhelm. Imputation strategies for clustering mixed-type data with missing values. Journal of Classification, 40(1):2–24, 2023.

[11] Donald B Rubin. Inference and missing data. Biometrika, 63(3):581–592, 1976.

[12] Marieke E Timmerman, Eva Ceulemans, Henk AL Kiers, and Maurizio Vichi. Factorial and reduced k-means reconsidered. Computational Statistics & Data Analysis, 54(7):1858–1871, 2010.

[13] R. Little and D. Rubin. Statistical Analysis with Missing Data. Wiley series in probability and statistics, New-York, 2002.

[14] Vincent Audigier, Ndèye Niang, and Matthieu Resche-Rigon. Clustering with missing data: which imputation model for which cluster analysis method? arXiv preprint arXiv:2106.04424, 2021.

[15] Julie Josse, François Husson, and Jérôme Pagès. Gestion des données manquantes en Analyse en Composantes Principales. Journal de la société française de statistique, 150(2):28–51, 2009.

[16] Yoshikazu Terada. Strong consistency of reduced k-means clustering. Scandinavian Journal of Statistics, 41(4):913–931, 2014.

# Sélection de modèles

# Une méthode de sélection de prédicteurs sous contrainte de non multicolinéarité dans les modèles linéaires généralisés

Christian Derquenne

*Chercheur indépendant - chris.emr@wanadoo.fr*

**Résumé.** Cet article se place dans le cadre de la multicolinéarité entre les prédicteurs au sein des modèles linéaires généralisés. Le phénomène de multicolinéarité peut entrainer des incohérences sur les coefficients de régression et des oublis de prédicteurs, cela peut par conséquent poser des problèmes d'interprétation qui peuvent entrainer de mauvaises décision. Le critère proposé est une nouvelle méthode de sélection de variables dans le cadre des modèles linéaires généralisés qui permet de respecter la non-multicolinéarité entre les prédicteurs. Ce critère est constitué de plusieurs statistiques : l'indépendance globale et marginale pour tester la non-multicolinéarité, l'ajustement global du modèle aux données, les effets marginaux des coefficients de régression multiple, la cohérence des signes de ceux-ci avec les coefficients de régression simple. Le modèle sélectionné possède deux propriétés : interprétabilité et prévision.

**Mots-clés.** Modèles linéaires généralisés, Multicolinéarité, méthodes de sélection de prédicteurs

**Abstract.** This article is placed within the framework of the multicollinearity between the predictors within a Generalized linear models The phenomenon of multicollinearity can lead to inconsistencies in the regression coefficients and omissions of predictors this can therefore pose problems of interpretation which can lead to poor decisions. The proposed criterion is a new method for selecting variables within the framework of generalized linear models which makes it possible to respect non-multicollinearity between the predictors. This criterion is made up of several statistics: the overall and marginal independence to test non-multicollinearity, the overall fit of the model to the data, the marginal effects of the multiple regression coefficients, the consistency of the signs of these with the coefficients simple regression. The selected model has two properties: interpretability and prediction.

**Keywords.** Generalized Linear model, Multicollinearity, stepwise methods

## 1 Contexte - objectif

La qualité des résultats d'un modèle statistique est primordiale pour les experts métier dans leurs domaines d'applications. Par exemple, en management de l'énergie, majoritairement l'objectif des modèles mis en oeuvre est d'obtenir la meilleure prévision de consommation d'électricité offrant les plus faibles erreurs (RMSE, MAPE, ...) afin de fournir des résultats fiables à destination de la production et donc de garantir l'équilibre offre/demande. Le modèle est très performant mais cela peut être au détriment de la compréhension des résultats en termes d'interprétabilité du sens et de l'influence des prédicteurs par les commanditaires de l'étude. Dans d'autres cas, l'objectif peut être justement de construire des modèles dans les résultats

sont interprétables et explicables. Et dans ces conditions, l'aspect qualité de la prévision n'est plus obligatoirement recherché. L'explicativité/interprétabilité des modèles, ou du moins de leurs résultats, devient par conséquent une discipline à part entière. Certaines approches statistiques sont plutôt dédiées à la prévision (réseaux de neurones, par exemple) et pour lesquelles il est nécessaire a posteriori de faire appel à des indicateurs numériques (LIME, valeurs de Shapley, cartes de salience, ...) si l'on désire interpréter les résultats. A l'opposé, les approches focalisées sur l'interprétabilité construisent des modèles permettant d'exploiter immédiatement le sens et l'importance des prédicteurs à l'aide par exemple des méthodes de régularisation, telles que les régressions Ridge, Lars [Hoerl et al., 1970], Lasso [Tibshirani et al., 1996] ou Elastic-Net [Zou et al., 2005] qui répondent en partie à la question de prédicteurs liés entre eux, mais elles exigent des paramètres non analytiques, donc difficiles à évaluer a priori. Le recours à la validation croisée est donc nécessaire. La régression sur composantes principales, qui consiste à réaliser une ACP sur toutes les variables explicatives, permet de garder au moyen de différentes stratégies certaines composantes principales qui résument au mieux les prédicteurs initiaux. Cependant, cette méthode n'est pas toujours optimale à l'égard du sens des coefficients et de l'importance des prédicteurs. La régression PLS (Partial Least Squares) [Wold H. et al., 1984] permet de mieux tenir compte de ces deux propriétés. En effet, cette méthode repose sur le principe de l'algorithme NIPALS (Non linear Iterative Partial Least Squares) qui est très puissant en termes de robustesse. Enfin, un premier critère nommé MCC (MultiCollinearity Criterium) a été proposé en 2022 [Derquenne, 2022] pour prendre en compte la multicolinéarité, dès le départ de la construction d'un modèle linéaire gaussien de régression multiple. Le critère MCC permet de maximixer les trois propriétés suivantes : reconstitution de l'inertie expliquée de l'estimateur MCO, adéquation des signes et de l'ordre des coefficients de régression multiple par rapport aux coefficients de corrélation linéaire simple associés. Une quatrième propriété a été introduite telle que la correspondance de la significativité des paramètres de régression multiple et simple. Un second critère nommé MCGC (MultiCollinearity Generalized Criterium) a permis d'étendre cette approche aux modèles linéaires généralisés [Derquenne, 2023]. Les trois premières propriétés à respecter dans le critère MCC sont adaptées au type de modèle, par exemple en régression logistique booléenne, une quatrième est rajoutée, telle que la proportion de bien classés.

Par ailleurs, la propriété de parcimonie d'un modèle est également très recherchée que ce soit dans un objectif de prévision ou dans un objectif d'interprétabilité. Le rapport qualité/prix de ces modèles est mesuré à l'aide de nombreux indicateurs, par exemple, les critères d'information AIC ou BIC. Les méthodes de sélection de prédicteurs permettent de garder les variables les plus influentes significativement dans le modèle, mais pas toujours de façon efficace face à la multicolinéarité qui met en défaut l'interprétabilité des résultats. En effet, comme déjà indiqué, les signes des coefficients de régression multiple peuvent être opposés à ceux des coefficients de corrélation linéaire simple, ou encore certains qui étaient fortement liés à la réponse peuvent être éliminés au détriment d'autres prédicteurs qui avaient à l'origine peu d'influence. Le modèle choisi peut alors être inintéressant sous l'aspect métier, voire incohérent et même dangereux à appliquer.

Le statisticien est alors face à un choix cornélien : soit construire un modèle statistique interprétable en éliminant la multicolinéarité quitte à garder des prédicteurs non significatifs ce

qui peut mettre en défaut la qualité d'ajustement sachant le prix payé par le nombre de variables explicatives, soit obtenir un modèle statistique parcimonieux et puissant avec des prédicteurs significatifs mais pénalisé par la multicolinéarité, donc inefficace en termes d'interprétabilité. Ce choix n'est heureusement pas binaire, en effet une solution potentielle serait un modèle puissant mais sans multicolinéarité. Dans ces conditions les résultats issus de ce modèle de régression multiple, notamment le sens de l'influence des prédicteurs et de leur importance pourraient être interprétés "en toute sécurité" car l'espace des variables candidates à l'explication sera structuré en séparant l'effet mutuel de chacune d'elles.

Dans ce article, nous proposons un nouveau critère permettant de construire un tel modèle statistique. La section 2 introduit ce nouveau critère nommé MCSRC (MultiCollinearity Stepwise Regression Criterium), puis il est appliqué à un exemple simulé utilisé dans [Derquenne, 2022] et comparé avec d'autres méthodes de sélection de prédicteurs. Dand la section 3, nous concluons sur les apports et les faiblesses de l'approche proposée, et nous fournissons quelques voies futures en termes d'amélioration et de nouveaux développements.

## 2 Un critère de sélection de prédicteurs en régression tenant compte de la multicolinéarité (MCSRC)

Rappelons que la multicolinéarité entre les prédicteurs d'un modèle de régression multiple peut entrainer d'une part, des signes contraires des coefficients de régression par rapport aux corrélations linéaires simples et d'autre part, des résultats de tests marginaux des coefficients (test $t$) en contradiction avec ce qui aurait pu être attendu d'après les tests sur les coefficients de régression simple. Pour pallier ces problèmes des solutions ont été proposées, comme indiqué dans l'introduction.

### 2.1 Le cas du modèle linéaire gaussien

Illustrons ce problème sur un jeu de données simulé. Ses caractéristiques sont les suivantes : $X_1 \rightarrow \mathcal{N}(0,1)$, $X_2 = X_1 + \mathcal{N}(0,1.96)$, $X_3 = 2X_2 + 3 + \mathcal{N}(0,0.04)$, $X_4 \rightarrow \mathcal{N}(0,2.25)$, $X_5 \rightarrow \mathcal{N}(0,1)$, $X_6 = X_5 + \mathcal{N}(0,0.04)$ et $Y = -1.5X_1 + 2X_2 + 0.5X_3 - 0.5X_4 + 0.1X_5 - 0.1X_6 + 4 + \mathcal{N}(0,0.25)$. Nous avons appliqué, (i) le critère classique des MCO, (ii) la régression sur composantes principales avec une sélection pas à pas de celles-ci reposant sur des tests statistiques de nullité des coefficients entrant et sortant, (iii) la régression sur premières composantes principales (RFPC) fondées sur une étape préalable de classification des prédicteurs [Derquenne et al., 2002], (iv) la régression PLS, (v) la sélection pas à pas des prédicteurs initiaux à l'aide des $p$-valeurs entrante et sortante (Stepwise Regression), (vi), la sélection ascendante de prédicteurs (Forward Regression), (vii) la méthode Incremental Forward Stagewise Regression [Hastie et al., 2007], (viii) le critère Lars, (ix) le critère Lasso, (x) le critère ElasticNet et (xi) le critère MCC proposé dans [Derquenne, 2022].

La table 1 fournit les coefficients de régression standardisés pour chacune des méthodes, ainsi que les corrélations simples. Premièrement, seuls RFPC et le multi-critère MCC fournissent des signes cohérents pour les six prédicteurs. Les forces de liaison sont relativement bien respectées pour la sélection pas à pas par $p$-valeurs, pour la régression PLS, avec un petit avantage pour le critère MCC. Seuls RFPC et MCC sont en parfaite adéquation avec l'ordre des coefficients

de régression et des corrélations simples. Globalement, les $R^2$ ajustés pour la plupart des méthodes sont supérieurs à 0,96, sauf pour RFPC (=0,66) et le multi-critère (=0,84). Ce dernier résultat est logique car MCC n'optimise pas seulement le critère des MCO. Enfin, *Eval* représente l'évaluation [Derquenne, 2022] de chaque méthode pa rapport à MCC. Ce dernier obtient "logiquement" la plus grande valeur (*Eval*=0,96), puis viennent Lasso, Lars, RFPC, ElasticNet, la régression sur composantes principales et PLS ($Eval \in [0,83; 0,85]$), enfin, MCO et les trois méthodes de sélection de prédicteurs ($Eval \in [0,73; 0,77]$). Par conséquent dans cet exemple, seul le critère MCC parait efficace face à la multicolinéarité en termes de reconstitution de force et de sens des liaisons tout en préservant la qualité du modèle. Ce résultat a été corroboré sur de nombreuses applications réelles et des simulations.

| | $\hat{\beta}^{MCO}$ | $\hat{\beta}^{PCR_1}$ | $\hat{\beta}^{RFPC}$ | $\hat{\beta}^{PLS}$ | $\hat{\beta}^{Spval}$ | $\hat{\beta}^{Forw}$ | $\hat{\beta}^{Swise}$ | $\hat{\beta}^{Lars}$ | $\hat{\beta}^{Lasso}$ | $\hat{\beta}^{ENet}_{\lambda=0,5}$ | $\hat{\beta}^{New}$ | $r_{yX}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | -0.379 | -0.381 | 0.246 | -0.323 | -0.383 | -0.383 | -0.380 | -0.263 | -0.263 | -0.209 | 0.083 | 0.171 |
| $X_2$ | 0.708 | 0.560 | 0.332 | 0.549 | 0.695 | 0.695 | 0.000 | 0.393 | 0.393 | 0.589 | 0.863 | 0.910 |
| $X_3$ | 0.419 | 0.569 | 0.332 | 0.557 | 0.435 | 0.435 | 1.130 | 0.608 | 0.608 | 0.594 | 0.866 | 0.913 |
| $X_4$ | -0.195 | -0.194 | -0.204 | -0.263 | -0.194 | -0.194 | -0.189 | -0.125 | -0.125 | -0.167 | -0.192 | -0.084 |
| $X_5$ | 0.099 | 0.064 | 0.041 | 0.051 | 0.051 | 0.051 | 0.000 | 0.000 | 0.000 | 0.071 | 0.046 | 0.136 |
| $X_6$ | 0.050 | 0.013 | -0.041 | 0.021 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | -0.008 | -0.083 |
| $R^2_{adj}$ | 0.987 | 0.987 | 0.663 | 0.980 | 0.987 | 0.987 | 0.981 | 0.962 | 0.962 | 0.953 | 0.840 | n.a |
| *Eval* | 0.786 | 0.830 | 0.850 | 0.827 | 0.774 | 0.774 | 0.727 | 0.854 | 0.854 | 0.830 | 0.958 | 1.000 |

Table 1: Coefficients de régression des méthodes

Intéressons-nous plus précisément aux résultats obtenus par les trois méthodes de sélection de prédicteurs : ascendante, pas à pas, et stagewise. La méthode de sélection ascendante consiste à entrer dans le modèle le prédicteur qui possède la plus petite $p$-valeur du coefficient de régression à condition qu'elle soit inférieure à un seuil de première espèce $\alpha_e$ fixé par le statiscien, par exemple 0,05, puis un deuxième prédicteur est ajouté au premier avec la condition précédente. L'algorithme se poursuit tant que la $p - valeur \leq \alpha_e$. Le problème de cette méthode est la non remise en cause des prédicteurs déjà entrés dans le modèle. La méthode de sélection pas à pas débute de la même façon que la précédente, mais elle a besoin de deux seuils $\alpha$'s, l'un pour les préditeurs entrants ($\alpha_e$) ; l'autre pour les prédicteurs sortants ($\alpha_s$). En effet, lors de l'ajout d'une variable ($p - valeur \leq \alpha_e$), si la $p$-valeur d'une autre déjà entrée dans le modèle dépasse le seuil $\alpha_s$, alors cette dernière est éliminée. L'algorithme continue tant que les deux seuils $\alpha$'s sont respectés. Enfin, la méthode stagewise ascendante crée un profil de coefficients comme suit : à chaque étape, il incrémente le coefficient de la variable la plus corrélée aux résidus courants d'une quantité $\pm\epsilon$, de signe déterminé par le signe de la corrélation. Efron et al. (2004) ont en fait considéré la version limitée de cet algorithme, avec $\epsilon \downarrow 0$, qui possède également des chemins de coefficients linéaires par morceau.

Comme nous pouvons le constater (cf. table 1) les méthodes de sélection ascendante et pas à pas au moyen du test de Student sur les coefficients fournissent les mêmes prédicteurs $X_1, X_2, X_3, X_4, X_5$, avec les $p$-valeurs respectives $< 0,0001$, 0,0002, 0,0154, $< 0,0001$ et $< 0,0001$, alors que l'approche stagewise sélectionne $X_1, X_3, X_4$ dont les trois $p$-valeurs sont toutes $< 0,0001$. Les

résultats montrent que les coefficients associés à $X_1$ sont négatifs pour les trois méthodes, alors que le coefficient de corrélation linéaire simple est positif. Sa $p$-valeur est égale à 0,0891, alors que le coefficient de régression multiple est très significatif. Comme indiqué, ce problème est typique de la multicolinéarité entre ces prédicteurs. Leur matrice de corrélations montre que certains sont fortement liés.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| $X_1$ | 1.000 (na) | 0.514 ($< 0,001$) | 0.506 ($< 0,001$) | 0.114 (0,261) | -0.031 (0.761) | -0.021 (0.826) |
| $X_2$ | 0.514 ($< 0,001$) | 1.000 (na) | 0.998 ($< 0,001$) | 0.136 (0,178) | 0.075 (0.460) | -0.044 (0.638) |
| $X_3$ | 0.506 ($< 0,001$) | 0.998 ($< 0,001$) | 1.000 (na) | 0.131 (0.195) | 0.069 (0.496) | -0.036 (0.692) |
| $X_4$ | 0.114 (0.261) | 0.136 (0.178) | 0.131 (0.195) | 1.000 (NA) | 0.047 (0.641) | -0.026 (0.755) |
| $X_5$ | -0.031 (0.761) | 0.075 (0.460) | 0.069 (0.496) | 0.047 (0.641) | 1.000 (na) | -0.970 ($< 0,001$) |
| $X_6$ | -0.021 (0.826) | -0.044 (0.638) | -0.036 (0.692) | -0.026 (0.755) | -0.970 ($< 0,001$) | 1.000 (na) |

Table 2: Matrice des coefficients de corrélation linéaire simple et $p$-valeurs associées

Le critère de sélection de prédicteurs proposé va tenir compte conjointement de la multicolinéarité, de la force de liaison entre les prédicteurs et la réponse, et de la cohérence des signes des coefficients de régression multiple avec les coefficients de corrélation simple.

### 2.1.1 Prise en compte de la multicolinéarité

Nous utilisons deux tests d'indépendance. Le premier est global ; le second est marginal.

***Test global d'indépendance*** : Soient $X_1, ..., X_j, ..., X_p$, $p$ variables gaussiennes de taille $n$ et soit $\mathbf{R}$, la matrice de corrélations linéaires de Pearson associée. Soit le jeu d'hypothèses $H_0$ : Indépendance linéaire entre les variables de $\mathbf{R}$ $vs$ $H_1$ : Il existe au moins une dépendance linéaire dans $\mathbf{R}$. La statistique de test est de la forme : $D = -(n-1)\log|\mathbf{R}|$ où $|\mathbf{R}|$ désigne le déterminant de $\mathbf{R}$, alors sous $H_0$, $D \to \chi^2_{p(p-1)/2}$

***Test marginal d'indépendance*** : Soient $X_1, ..., X_j, ..., X_p$, $p$ variables gaussiennes de taille $n$ et soit $R_j^2$, le coefficient de détermination du modèle linéaire : $X_j^* = \sum_{k \neq j} \gamma_k X_k + \epsilon$. $R_j^2$ sera d'autant plus élevé que $X_j$ sera corrélé avec des variables $X_k$'s. Soit le jeu d'hypothèses $H_0$ : $X_j$ est indépendante linéairement des $p-1$ autres variables $vs$ $H_1$ : Il existe au moins une variable parmi les $p-1$ autres liée linéairement avec $X_j$. La statistique de test est de la forme : $F_j = [R_j^2(n-p-2)]/[(1-R_j^2)(p-1)]$, alors sous $H_0$ $F_j \to \mathcal{F}(p-1, n-p-2)$

### 2.1.2 Force de liaison des prédicteurs avec la réponse

Nous considérons le modèle de régression linéaire multiple suivant : $Y = \mathbf{X}\beta + \epsilon$ où $\epsilon \to \mathcal{N}(0, \sigma^2)$

Comme pour la multicolinéarité, nous utilisons deux tests global et marginal.

***Test global d'ajustement*** : Soit $Y$, la réponse et soient $X_1, ..., X_j, ..., X_p$, les $p$ prédicteurs numériques de taille $n$. Soit le jeu d'hypothèses $H_0$ : $\beta_1 = ... = \beta_j = ... = \beta_p = 0$ $vs$ $H_1$ : Il existe un $\beta_j \neq 0$. La statistique de test est de la forme : $F = [R^2(n-p-1)]/[(1-R^2)p]$, où $R^2$ est le coefficient de détermination global du modèle, alors sous $H_0$ $F \to \mathcal{F}(p, n-p-1)$.

***Test marginal d'ajustement*** : Soit $Y$, la réponse et soient $X_1, ..., X_j, ..., X_p$, les $p$ prédicteurs de taille $n$. Soit le jeu d'hypothèses $H_0 : \beta_j = 0$ *vs* $H_1 : \beta_j \neq 0$. La statistique de test est de la forme : $t(X_j) = \beta_j/\sigma_{\beta_j}$, où $\sigma_{\beta_j}$ est l'écart-type du coefficient $\beta_j$, alors sous $H_0$, $t(X_j) \to t_{n-p-1}$.

### 2.1.3 Forme du critère MCSRC

Le critère MCSRC est composé des résultats des quatre tests précédents et, de la cohérence des signes entre les coefficients de régression et les coefficients de corrélation simple. Afin de décider si un modèle peut être sélectionné, il est nécessaire de fixer des seuils d'acceptation. Ceux-ci sont obtenus à l'aide des risques de première espèce des tests proposés précédemment comme dans les méthodes classiques de sélection ascendante, descendante et pas à pas.

Pour les deux tests d'indépendance, garder l'hypothèse nulle correspondra à la non-multicolinéarité et donc l'objectif recherché, alors que pour les deux tests sur la force des liaisons des prédicteurs, le but recherché sera le rejet de l'hypothèse nulle. Formalisons ces voeux.

Soient $\alpha_D$, $\alpha_r$, $\alpha_F$ et $\alpha_t$, les risques de première espèce choisis pour les tests d'indépendance globale, marginale, d'ajustement global et marginal et soient $pval_D$, $pval_r(j)$, $pval_F$ et $pval_t(j)$, les $p$-valeurs obtenues sur ces tests. Signalons que $pval_r(j)$ et $pval_t(j)$ correspondent aux $p$-valeurs spécifiques à la variable $X_j$. Pour les deux tests d'indépendance, l'acceptation correspond à $pval_D \geq \alpha_D$ et $pval_r(j) \geq \alpha_r$, alors que pour les deux tests d'ajustement nous avons : $pval_F \leq \alpha_F$ et $pval_t(j) \leq \alpha_t$. Signalons que le test marginal d'indépendance n'est pas utilisé dans le cas $p = 2$.

Le critère MSC pour le modèle linéaire gaussien prend la forme suivante :

$$A_1 = \left[ \left( \frac{pval_D - \alpha_D}{1-\alpha_I} + \frac{1}{p} \sum_{j=1}^{p} \frac{pval_r(j) - \alpha_r}{1-\alpha_r} + \frac{\alpha_F - pval_F}{\alpha_F} + \frac{1}{p} \sum_{j=1}^{p} \frac{\alpha_t - pval_t(j)}{\alpha_t} \right) \Big/ 4 \right] \quad (1)$$

$$A_2 = 1_{[pval_D \geq \alpha_D]} \times \prod_{j=1}^{p} 1_{[pval_r(j) \geq \alpha_r]} \times 1_{[pval_F \leq \alpha_F]} \times \prod_{j=1}^{p} 1_{[pval_t(j) \leq \alpha_t]} \times \prod_{j=1}^{p} 1_{[\beta_j \times r_{yX_j} \geq 0]} \quad (2)$$

Enfin $C_{MCSRC}^{(MLG)} = A_1 \times A_2 \in [0, 1]$. Pour que le critère soit strictement positif, il est nécessaire que les $3p + 2$ contraintes dans $A_2$ soient respectées. L'objet $A_1$ permet de moduler la qualité des modèles qui respectent la multicolinéarité conjointement et la force d'ajustement du modèle. Par conséquent le vecteur des coefficients de régression du modèle sélectionné sera :

$$\tilde{\beta} = \arg \max_{m \in \mathcal{M}} C_{MCSRC}^{(MLG)}(m) \quad (3)$$

où $m$ est un modèle appartenant à $\mathcal{M}$, l'ensemble des $2^p - 1$ modèles possibles.

Un autre intérêt de ce critère est de pouvoir calculer la contribution de chaque prédicteur à la part d'inertie expliquée par le modèle. Celle-ci est fournie par le coefficient de détermination $R^2$ qui peut être mis sous la forme suivante : $\sum_{j=1}^{p} \tilde{\beta}_j r_{yX_j}$, où $\tilde{\beta}_j$ est le coefficient standardisé de régression multiple associé à $X_j$, alors le pourcentage de contribution de chaque prédicteur est donnée par :

$$CTR(X_j) = \frac{\tilde{\beta}_j r_{yX_j}}{\sum_{j=1}^{p} \tilde{\beta}_j r_{yX_j}} \times 100\% \quad (4)$$

***Remarque*** : Cet indicateur est seulement utilisable si les signes du coefficient de régression multiple et de corrélation linéaire simple sont identiques (possible absence de multicolinéarité).

Appliquons le critère proposé à l'exemple simulé. Nous avons choisi les risques de première espèce suivants : $\alpha_D = 0,01$, $\alpha_F = 0,01$, $\alpha_r = 0,01$ et $\alpha_t = 0,025$. Les prédicteurs sélectionnés sont $X_3$, $X_4$ et $X_5$. Nous pouvons constater sur la table 3 que les coefficients estimés de $X_1$ par les méthodes de sélection pas à pas, ascendante et stagewise sont négatifs et significatifs, alors que la corrélation simple est positive et non significative. Le critère MCSRC ne retient pas $X_1$. $X_2$ n'est pas sélectionné par stagewise et MCSRC, alors qu'elle l'est par les deux autres méthodes. Par contre $X_3$ est tout le temps choisi, mais avec une plus faible $p$-valeur pour stagewise et MCSRC. Ces deux derniers résultats sont logiques car $X_2$ et $X_3$ sont très fortement liés (cf. table 2). La variable $X_4$ est retenue par les quatre méthodes avec de très faibles $p$-valeurs alors que la corrélation simple avec la réponse est non significative. $X_5$ est toujours sélectionnée sauf par stagewise, ce qui est logique car la corrélation simple n'est pas significative. Et $X_6$ n'est jamais choisie, ce qui est logique car la corrélation simple n'est pas significative, non plus.

Seul le critère MCSRC respecte la non multicolinéarité, la valeur du critère MCSRC est de 0,856, alors qu'elle est nulle pour tous les autres. Remarquons également que le $R^2_{adj} = 0,879$ pour MCSRC est relativement proche de ceux des autres critères (entre 0,981 et 0,987) : peu de perte de qualité d'ajustement et respect de la non-multicolinéarité. De plus, la valeur obtenue pour le critère MCC [Derquenne, 2022] construit pour le modèle linéaire gaussien qui vaut 0,833 (ligne *Eval* du tableau 3) est supérieure à celles des autres méthodes. Rappelons que ce critère est dans $[0, 1]$, plus une valeur est proche de 1, plus le modèle obtenu est optimal face à la multicolinéarité. Lors de la phase de régularisation de la matrice de corrélations entre les prédicteurs, le critère MCC fournit une valeur optimale $\tilde{\delta}$ pour éliminer la multicolinéarité. Dans le cas du modèle sélectionné par le critère MCSRC ($X_3$,$X_4$,$X_5$), nous obtenons $\tilde{\delta} = 0,131$. Cela signifie que les corrélations entre ces trois prédicteurs sont inférieures à 0,131, en valeur absolue.

Signalons enfin que la méthode proposée peut fournir aussi un ensemble de modèles candidats acceptables (la valeur du critère est stictement positive). Dans cet exemple, les modèles potentiels sont par ordre décroissant du critère : $\{X_2, X_4, X_5\}$ (MSC=0,835,$R^2_{adj} = 0,873$) ; $\{X_3, X_4\}$ (MSC=0,796,$R^2_{adj} = 0,873$) ; $\{X_2, X_4\}$ (MSC=0,792,$R^2_{adj} = 0,869$) ; $\{X_3, X_5\}$ (MSC=0,775,$R^2_{adj} = 0,836$). Tous ces modèles respectent la cohérence des signes de coefficients de régression multiple et de corrélation simple.

| | $\hat{\beta}^{MCO}$ | $\hat{\beta}^{Spval}$ | $\hat{\beta}^{Forw}$ | $\hat{\beta}^{Swise}$ | $\hat{\beta}^{MSC}$ | $r_{y_X}$ |
|---|---|---|---|---|---|---|
| $X_1$ | -0.379 (< 0.001) | -0.383 < 0.001) | -0.383 (< 0.001) | -0.380 (< 0.001) | 0.000 (na) | 0.171 (0.089) |
| $X_2$ | 0.708 (< 0.001) | 0.695 (< 0.001) | 0.695 (< 0.001) | 0.000 (na) | 0.000 (na) | 0.910 (< 0.001) |
| $X_3$ | 0.419 (0.020) | 0.435 (0.015) | 0.435 (0.015) | 1.130(< 0.001) | 0.935 (< 0.001) | 0.913 (< 0.001) |
| $X_4$ | -0.195 (< 0.001) | -0.194 (< 0.001) | -0.194 (< 0.001) | -0.189 (< 0.001) | -0.210 (< 0.001) | -0.084 (0.406) |
| $X_5$ | 0.099 (0.049) | 0.051 (< 0.001) | 0.051 (< 0.001) | 0.000 (na) | 0.046 (0.023) | 0.081 (0.179) |
| $X_6$ | 0.050 (0.320) | 0.000 (na) | 0.000 (na) | 0.000 (na) | 0.000 (na) | -0.083 (0.414) |
| $R^2_{adj}$ | 0.987 | 0.987 | 0.987 | 0.981 | 0.879 | n.a |
| *Eval* | 0.786 | 0.774 | 0.774 | 0.727 | 0.833 | 1.000 |

Table 3: Coefficients de régression des méthodes de sélection de prédicteurs

### 2.2 Adaptation du critère MCSRC aux modèles linéaires généralisés

Nous nous restreindrons au modèle logit booléen et au modèle logit polytomique ordonné. Pour chacun d'eux, le critère est modifié par l'ajout de nouvelles conditions. Dans le premier modèle, la statistique de Fisher est remplacée par le rapport de vraisemblances estimé et par la déviance permettant de juger l'adéquation du modèle aux données ; le second tient compte de ces deux éléments, ainsi que de la propriété du rapport de côtes proportionnelles.

#### 2.2.1 Modèle logit booléen

Dans le cadre de la régression logistique binaire [Derquenne, 2023], l'étude de la multicolinéarité a montré que celle-ci entrainait les mêmes problèmes pour les signes des coefficients et pour les tests marginaux associés. Les méthodes de sélection de préditeurs subissent par conséquent les mêmes écueils.

Soit $Y \in \{0;1\}$, la réponse booléenne et soient $X = (X_1, ..., X_j, ..., X_p)$, les $p$ prédicteurs numériques de taille $n$. Nous considérons le modèle de régression logistique multiple dichotomique : $Pr[Y = 1/X] = \frac{exp^{\beta_0 + X\beta}}{1+exp^{\beta_0+X\beta}}$, où $\beta$ est le vecteur des $p$ coefficients. Les tests sur la force de liaison des prédicteurs et sur l'adéquation du modèle aux données sont les suivants.

***Test du rapport de vraisemblances***, soit le jeu d'hypothèses $H_0 : \beta_1 = ... = \beta_j = ... = \beta_p = 0$ *vs* $H_1$ : Il existe un $\beta_j \neq 0$. La statistique de test est de la forme : $LR = -2(l_{H_0} - l_{est})$, où $l_{H_0}$ et $l_{est}$ sont respectivement la log-vraisemblance sous l'hypothèse nulle et la log-vraisemblance estimée, alors sous $H_0$, $LR \rightarrow \chi^2_p$ où $p$ est le nombre de prédicteurs. $\alpha_{LR}$ et $pval_{LR}$ sont respectivement le risque de première espèce et la $p$-valeur du test.

***Test de la déviance***, soit le jeu d'hypothèses $H_0$ : adéquation du modèle aux données *vs* $H_1$ : non adéquation. La statistique de test est de la forme $Dev = -2(l_{est} - l_{sat})$, où $l_{sat}$ est la log-vraisemblance saturée, alors sous $H_0$, $Dev \rightarrow \chi^2_{n-p-1}$ où $n$ est le nombre d'observations. $\alpha_{Dev}$ et $pval_{Dev}$ sont respectivement le risque de première espèce et la $p$-valeur du test.

Les deux statistiques pour évaluer la multicolinéarité sont celles utilisées pour le modèle linéaire gausienne, cela se justifie par le fait que la matrice de corrélations $\mathbf{R}$ entre les prédicteurs numériques se retrouve dans l'estimateur du maximum de vraisemblance et influence par conséquent les coefficients et les tests associés à ceux-ci [Derquenne, 2023]. Enfin, les coefficients de régression sont testés à l'aide de la statistique de Wald, analogue à celle de Student du modèle linéaire gaussien. $\alpha_w$ et $pval_w(j)$ sont respectivement le risque de première espèce et la $p$-valeur pour le prédicteur $X_j$. Signalons que le test marginal d'indépendance n'est pas utilisé dans le cas $p = 2$.

***Le critère MCSRC pour le modèle logit dichotomique*** est de la forme suivante :

$$B_1 = \left[\left(\frac{pval_D - \alpha_D}{1-\alpha_D} + \left(\frac{1}{p}\sum_{j=1}^{p}\frac{pval_r(j)-\alpha_r}{1-\alpha_r} + \frac{\alpha_w - pval_w(j)}{\alpha_w}\right) + \frac{\alpha_{LR} - pval_{LR}}{\alpha_{LR}} + \frac{pval_{Dev}-\alpha_{Dev}}{1-\alpha_{Dev}}\right)\Big/ 5\right] \quad (5)$$

$$B_2 = 1_{[pval_D \geq \alpha_D]} \times \prod_{j=1}^{p} 1_{[pval_r(j) \geq \alpha_r]} \times 1_{[pval_{LR} \leq \alpha_{LR}]} \times \prod_{j=1}^{p} 1_{[pval_w(j) \leq \alpha_w]} \times 1_{[pval_{Dev} \geq \alpha_{Dev}]} \times \prod_{j=1}^{p} 1_{[\beta_j^{mul} \times \beta_j^{uni} \geq 0]} \quad (6)$$

où $\beta_j^{mul}$ et $\beta_j^{uni}$ sont respectivement les coefficients de régression logistique multiple et simple. Enfin $C_{MCSRC}^{(logit(0,1))} = B_1 \times B_2 \in [0,1]$. Pour que le critère soit strictement positif, il est nécessaire

que les $3p+3$ contraintes dans $B_2$ soient respectées. L'objet $B_1$ permet de moduler la qualité des modèles qui respectent la non-multicolinéarité conjointement à la force d'ajustement du modèle et l'adéquation de celui-ci aux données. Par conséquent le vecteur des coefficients de régression du modèle sélectionné sera :

$$\tilde{\beta} = \arg\max_{m \in \mathcal{M}} C_{MCSRC}^{(logit(0,1))}(m) \tag{7}$$

### 2.2.2 Modèle logit ordinal

Soit $Y \in \{1; ...; k; ...; ...; r\}$, la réponse ordinale à $r$ catégories ordonnées et soient $X = (X_1, ..., X_j, ..., X_p)$, les $p$ prédicteurs numériques de taille $n$. Nous considérons le modèle de régression logistique multiple ordinale : $Pr[Y \leq 1/X] = \frac{exp^{\theta_1 + X\beta_1}}{1+exp^{\theta_1+X\beta_1}}$, ..., $Pr[Y \leq k/X] = \frac{exp^{\theta_k+X\beta_k}}{1+exp^{\theta_k+X\beta_k}}$, ..., $Pr[Y \leq r-1/X] = \frac{exp^{\theta_{r-1}+X\beta_{r-1}}}{1+exp^{\theta_{r-1}+X\beta_{r-1}}}$. Ce modèle est nommé modèle logit à rapport de côtes proportionnelles, si les $r-1$ vecteurs de coefficients $\beta_k$, pour $k = 1, 2, ..., r-1$ sont égaux. En d'autres termes, $\forall, j = 1, ..., p; \beta_{1,j} = ... = \beta_{k,j} =, ... = \beta_{r-1,j}$. Cette condition représente l'hypothèse nulle du test du rapport de côtes proportionnelles (proportionnal odds ratio).

Deux types de tests peuvent être réalisés : le premier est global car il compare les $r-1$ vecteurs de coefficients, cela revient à tester si les $r-1$ pentes sont parallèles, il y a dans ce cas $(r-2)p$ égalités à vérifier ; le second teste marginalement pour chaque préditeur $X_j$, si $\beta_{1,j} = ... = \beta_{k,j} = , ... = \beta_{r-1,j}$, il n'y a dans ce cas que $r-2$ égalités à respecter.

***Le test global du rapport de côtes proportionnelles*** peut être effectué à l'aide des statistiques du rapport de vraisemblances, du score et de Wald. Par exemple, la statistique du rapport de vraisemblances prend la forme suivante : $LR_{OR} = -2(l_{orp} - l_{norp})$, où $l_{orp}$ et $l_{norp}$ sont respectivement la log-vraisemblance du modèle à rapport de côtes proportionnelles et la log-vraisemblance du modèle général. Sous $H_0$, $LR_{OR} \rightarrow \chi^2_{(r-2)p}$. $\alpha_{OR}$ et $pval_{OR}$ sont respectivement le risque de première espèce et la $p$-valeur du test.

***Le test marginal*** peut être effectué notamment à l'aide du test de Brant [Brant, 1990]. Pour chaque prédicteur, Brant teste sous $H_0$, l'égalité des coefficients adjacents, tel que : $\beta_{\leq k,j} = \beta_{>k:r,j}$ pour $k = 1, r-1$ où $\beta_{\leq k,j}$ et $\beta_{>k:r,j}$ sont respectivement les coefficients univariés des modèles : logit$(Y \leq k \ vs \ Y > k)$ et logit$(Y \leq k+1 \ vs \ Y > k+1)$. La statistique de Wald prend la forme suivante : $w_{jk} = (\beta_{\leq k,j} - \beta_{>k:r,j})^2 / (\sigma^2_{\beta_{\leq k,j}} + \sigma^2_{\beta_{>k:r,j}})$. Sous $H_0$, $w_{jk} \rightarrow \chi^2_1$, lorsque $r = 3$. Si $r > 3$, alors cette statistique est composée de $r-2$ éléments et elle suivra un $\chi^2_{r-2}$. $\alpha_B$ et $pval_{B(j)}$ sont respectivement le risque de première espèce et la $p$-valeur du test. Signalons que ce test marginal n'est pas utilisé dans le cas $p = 2$.

***Le critère MCSRC pour le modèle logit ordinal*** est de la forme suivante :

$$C_1 = \left[ \left( 5B_1 + \frac{pval_{OR} - \alpha_{OR}}{1-\alpha_{OR}} + \frac{1}{p} \sum_{j=1}^{p} \frac{pval_B(j) - \alpha_B}{1-\alpha_B} \right) \Big/ 7 \right] \tag{8}$$

$$C_2 = B_1 \times 1_{[pval_{OR} \geq \alpha_{OR}]} \times \prod_{j=1}^{p} 1_{[pval_B(j) \geq \alpha_B]} \tag{9}$$

Enfin $C_{MCSRC}^{(logit(ord))} = C_1 \times C_2 \in [0, 1]$. Pour que le critère soit strictement positif, il est nécessaire que les $4p+4$ contraintes dans $C_2$ soient respectées. L'objet $C_1$ permet de moduler la qualité des modèles qui respectent la non-multicolinéarité, la force d'ajustement du modèle, l'adéquation

de celui-ci aux données et le rapport de côtes proportionnelles. Le vecteur des coefficients de régression du modèle sélectionné sera :

$$\tilde{\beta} = \arg\max_{m \in \mathcal{M}} C_{MCSRC}^{(logit(ord))}(m) \tag{10}$$

## 3 Apports, applications et voies futures

Le critère MCSRC proposé est une nouvelle méthode de sélection de variables dans le cadre des modèles linéaires généralisés qui permet de respecter la non-multicolinéarité entre les prédicteurs. Ce critère est constitué de plusieurs statistiques : l'indépendance globale et marginale pour tester la non-multicolinéarité, l'ajustement global du modèle aux données, les effets marginaux des coefficients de régression multiple, la cohérence des signes de ceux-ci avec les coefficients de régression simple, et dans le cas du modèle logit ordinal, le rapport de côtes proportionnelles. L'avantage de MCSRC sur les autres méthodes de sélection de variables réside en particulier sur la possibilité d'interpréter les résultats du modèle en "toute sécurité" car l'effet néfaste de la multicolinéarité ne peut pas apparaitre dans le modèle sélectionné. Cet avantage est renforcé par la préservation de la qualité d'ajustement du modèle pour expliquer la réponse qui reste comparable aux autres méthodes. En d'autres termes, le modèle sélectionné possède deux propriétés : interprétabilité et prévision. Le critère introduit dans cet article a été évalué sur différents cas simulés (pas de corrélation entre prédicteurs, corrélations modérées, très fortes corrélations), ainsi que sur d'autres exemples issus de la littérature, et des cas réels d'applications dans les domaines marketing, médical et management de l'énergie. Les voies futures vont consister à établir les propriétés mathématiques du critère proposé, l'étendre à la prévision de chroniques à l'aide de plusieurs prédicteurs temporels et à la modélisation multivariée de réponses.

## Bibliographie


Bastien Ph., Esposito Vinzi E. & Tenenhaus M., (2005), PLS generalized linear regression, *Computational Statistics & Data Analysis*, **48**(1), 17-46.

Brant R., (1990), Assessing proportionality in the proportional odds model for ordinal logistic regression, *Biometrics* **46**, 1171-1178.

Derquenne Ch., (2022), Un multi-critère pour contrôler la multicolinéarité dans les modèles linéaires de régression multiple, *53ièmes Journées de Statistique*, Lyon, France, 538-543.

Derquenne Ch., (2023), Un multi-critère pour traiter la multicolinéarité dans les modèles linéaires généralisés, *54ièmes Journées de Statistique*, Bruxelles, Belgique.

Hoerl, A.E. & Kennard R.W., (1970), Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 42 (1)*, 80-86.

Tibshirani, R., (1996), Regression Shrinkage and Selection via the Lasso, *J. R. Statist. Soc. B*, **58**, *No.* 1, 267-288.

Wold S., Ruhe A., Wold H. & Dunn III W.J., (1984), The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM J. Sci. Stat. Comput.*, **5**, n°3, 735-743.

Zou H. & Hastie T., (2005), Regularization and Variable Selection via the Elastic Net, *J. R. Statist. Soc. B*, **67**, *No.* 2, 301-320.

# MODEL SELECTION FOR CONTEXTUAL BANDITS

Julien Aubert [1] & Luc Lehéricy [2] & Patricia Reynaud-Bouret [3]

[1] *Laboratoire J.A. Dieudonné, Université Côte d'Azur, France - jaubert@unice.fr*
[2] *Laboratoire J.A. Dieudonné, Université Côte d'Azur, France - luc.lehericy@univ-cotedazur.fr*
[3] *Laboratoire J.A. Dieudonné, Université Côte d'Azur, France - patricia.reynaud-bouret@univ-cotedazur.fr*

**Résumé.** Ce travail aborde le problème de la sélection de modèles dans les algorithmes de bandits contextuels lorsqu'ils sont utilisés pour modéliser une tâche d'apprentissage. Plus précisément, chaque modèle représente une partition de l'espace des contextes sur chaque ensemble de laquelle un algorithme de bandit est appliqué. Notre objectif est de trouver le modèle qui correspond le mieux aux données d'apprentissage. En étendant les outils traditionnels d'estimation et de sélection de modèle aux données non i.i.d et non stationnaires, nous montrons dans un premier temps qu'une procédure de hold-out sur les données satisfait un taux de convergence classique. Ensuite, sous diverses hypothèses, nous formulons des inégalités oracles avec différents taux de convergence. Nous fournissons également des exemples pour lesquels les hypothèses sont satisfaites. Enfin, nous testons nos résultats sur des données d'apprentissage synthétiques et réelles.

**Mots-clés.** Estimation statistique, Sélection de modèle, Bandits contextuels, Cognition, Vraisemblance pénalisée, Hold-out.

**Abstract.** This work tackles the problem of model selection in contextual bandit algorithms when used to model a learning task. More specifically, each model represents a partition of the contexts space on each set of which a bandit algorithm is applied. Our goal is to find the model that best fits the learning data. By extending the traditional tools of estimation and model selection to non i.i.d and non stationary data, we show first that a hold out procedure on the data satisfies a classical rate of convergence. Then, under various assumptions, we formulate oracle inequalities with various rates of convergence. We also provide examples for which the assumptions are satisfied. Finally, we test our results on both synthetic and real learning data.

**Keywords.** Satistical estimation, Model selection, Contextual bandits, Cognition, Penalized likelihood, Hold-out.

# 1 Introduction

## 1.1 Cognitive Models

Imagine an individual learning a categorisation task. Without initially knowing the rule, the individual is presented with objects in a sequential way, each object having a certain amount of features, and has to classify the objects in two categories. The reward he gets is whether the object was classified in the right category. His goal is to identify what the rule for belonging to one category or the other is. Our goal is to estimate its learning strategy, meaning to understand which features of the objects the learner saw as important for belonging to each of the categories.

A model describing the previous experiment is called a categorization model. It belongs to the wider category of cognitive models. They help understanding the mechanisms that occur in the brain while learning, remembering, and predicting tasks. They have been widely studied in the cognition literature [1] and have a major impact on education for instance. In this article, we consider cognitive models that describe learning experiments. Back to our example, the choice made by the learner to classify an object in a category depends on the choices he made before.

To assess the relevance of a model, one should first evaluate its goodness of fit through maximum likelihood estimation (MLE) or least-square regression. Then, one should compare the given model to other models and choose the best model given a certain criterion. This empirical methodology is traditionally used for cognitive modelling [2]. However, no theoretical result exists to prove the relevance of either the model or the methodology. Our goal is to establish a theoretical framework for model selection in learning.

## 1.2 Contextual Bandits

To do so, we use machine learning (ML) algorithms. We try to infer the learning strategy of a ML model. More specifically, given learning data, our goal is to select the learning algorithm that best fits the learning data. The algorithms we work with are contextual bandits algorithms (see 1). For $d \in \mathbb{N}$, let $\mathcal{P}_d$ be the probability simplex $\{q \in [0,1]^{d+1}, \|q\|_1 = 1\}$. Let $K \geqslant 2$ be the number of actions, and let $\mathcal{X}$ be a set representing the context space. We are interested in estimating the parameters of adversarial contextual bandits algorithms which satisfy, by definition, the following interaction protocol with the environment

**Algorithm 1** Interaction protocol for contextual bandits [3]

Adversary secretly chooses a sequence of losses (or rewards) $(\pi_t)_{t=1}^n$ with $\pi_t \in [0,1]^K$.
Adversary secretly chooses a sequence of contexts $(x_t)_{t=1}^n$ with $x_t \in \mathcal{X}$.
**for** $t = 1, 2, \ldots, n$ **do**
    Learner observes context $x_t \in \mathcal{X}$
    Learner selects a distribution $p_t \in \mathcal{P}_{K-1}$ and samples $A_t$ from $p_t$.
    Learner observe loss $\pi_{A_t,t}$.

In this setting a policy is a function mapping history sequences to distributions over actions. The regret measures the performance of a policy by comparing the rewards collected by the learner and the rewards collected by the best policy. Many algorithms are able to cope with problem 1. Contextual bandits refer to the class of bandit problems in which the learner has access to additional information (context) at each time step. There are many applications of such algorithms such as ad recommendation, patient follow-up in healthcare, etc... (see [4] for more details)

When the context space $\mathcal{X}$ is small enough, a simple idea to solve the contextual bandit problem is to apply one bandit algorithm per context. By doing so, one can show that the regret is lower bounded by $2\sqrt{nK|\mathcal{X}|\log(K)}$. If the context space is too large or if the amount of data is too small, then applying one bandit per context will always be insufficient. If instead of considering the set of all policies from $\mathcal{X}$ to $[K]$, one considers the set of all policies that are constant on each part of a given partition $\mathcal{P}$ of $\mathcal{X}$, then one can apply a Bandit algorithm on each part of the partition and therefore the factor $|\mathcal{X}|$ becomes $|\mathcal{P}|$ in the regret. This is the framework we will be using.

## 1.3 Link between Contextual Bandits and Cognitive Models

If we go back to our categorization example, at each time step, the learner has access to the features of each object. We make the assumption that he is learning by rule [1]: an object is in one category or the other because it obeys some criteria. The learner is thus partitioning the set of objects into parts each representing a different rule. On each part, we assume that the learner is applying a bandit algorithm. We therefore are in the setting of contextual bandits, only considering the set of policies that are constant on each part of a partition of the set of contexts $\mathcal{X}$. For the categorization task, the contexts are the objects the learner is seeing at each time step.

Our goal is to estimate the way the learner partitions the set of objects he has to classify. In other words, given that he is learning by making a partition on the set of contexts and applying a bandit algorithm in each set of the partition, can we estimate the partition he actually used to learn ? Each model we study is thus a partition of the context space on each set of which a bandit algorithm is applied.

From a statistical point of view, the underlying problem is similar to selecting the best model (i.e. closest to reality) when constructing a histogram [5]. Unlike the traditional framework in which the data are i.i.d with a common density to estimate, here the learning data are not stationary and not independent. During a learning task, present choices should be affected by past choices and rewards. Thus, traditional results and tools [5] such as cross validation [6] would not apply here. In a previous work, [7] focus on estimating the parameters of Exp3 when fitting it to learning data. In this work, we tackle the problem of choosing the best model to fit the learning data.

We extend the work of [5] for model selection to non i.i.d and non stationary data. To do so, we use classical concentration inequalities for martingales [8].

### 1.4 Contributions

- We show that a hold out procedure in this framework satisfies a classical rate of convergence, regardless of how the initial parameters are estimated.
- Let $\mathcal{M}$ be a collection of models. Let $\hat{\theta}^m = \arg\max_{\theta^m} \ell_T(\theta^m)$ be the estimated parameter for model $m \in \mathcal{M}$, where $\ell_T(\theta^m)$ is the log-likelihood function stopped at time $T$. Let $\text{pen} : \mathcal{M} \to \mathbb{R}_+$ be a penalty function. Let $\hat{m}$ be the model that minimize the penalized criterion:
$$\widehat{m} := \arg\min_{m \in \mathcal{M}} \left( -\frac{\ell_T(\hat{\theta}^m)}{T} + \text{pen}(m) \right)$$
Then under some assumptions, for a well chosen $T$, we show that if the penalty function scales as $1/T$, then the estimated distribution satisfies an oracle inequality with a $O(1/T)$ additive factor.
- We provide examples with Stochastic Gradient Bandits [9] and EXP3-IX [3] for which the assumptions are satisfied.
- Finally, we test our results on both synthetic and real learning data on a categorization task [1].

# References


[1] G. Mezzadri. *Statistical inference for categorization models and presentation order*. Phd thesis, Université Côte d'Azur, LJAD, France, 2020.

[2] R. C. Wilson and A. G.E. Collins. Ten simple rules for the computational modeling of behavioral data. *Elife*, 8:e49547, 2019.

[3] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[4] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits, 2019.

[5] Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[6] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it?, 2022.

[7] Julien Aubert, Luc Lehéricy, and Patricia Reynaud-Bouret. On the convergence of the MLE as an estimator of the learning rate in the exp3 algorithm. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1244–1275. PMLR, 7 2023.

[8] Y. Baraud. A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. *Bernoulli*, 16(4):1064 – 1085, 2010.

[9] Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *International Conference on Machine Learning*, pages 24325–24360. PMLR, 2023.

# Mutant-UCB : entre bandits et algorithme évolutionnaire, une approche pour la sélection de modèles

Julie Keisler[1] & Margaux Brégère[2]

[1] *EDF R&D, INRIA Lille Nord Europe - France - julie.keisler@edf.fr*
[2] *EDF R&D, LPSM Sorbonne Université - France - margaux.bregere@edf.fr*

**Résumé.** Nous présentons la sélection de modèles comme un problème de bandits avec un nombre infini de bras (*infinite-armed bandit*). Les modèles sont les bras du problème de bandit sous-jacent, et le choix d'un bras correspond à un entraînement partiel du modèle (allocation de ressources). La récompense est la performance du modèle sélectionné après son entraînement partiel. Sélectionner le meilleur modèle revient à identifier le meilleur bras (*best-arm identification*). On définit alors le regret comme l'écart entre la performance du modèle optimal et celle du modèle finalement choisi. En partant de cette modélisation, nous proposons un nouvel algorithme, appelé Mutant-UCB, qui généralise l'algorithme UCB-E (développé par Audibert et al. 2010) au problème de bandit stochastique avec un nombre de bras infini, et y incorpore des opérateurs des algorithmes évolutionnaires. Nous avons testé cet algorithme pour optimiser des réseaux de neurones (architectures et hyperparamètres) sur trois jeux de données de classification d'images. Notre algorithme se révèle plus performant que l'état de l'art sur chacun de ces jeux de données ce qui montre que cette approche hybride est pertinente.

**Mots-clés.** Bandits avec une infinité de bras, sélection de modèles, optimisation d'architectures, optimisation d'hyperparamètres, algorithme évolutionnaire, classification d'images, AutoML, apprentissage en ligne

**Abstract.** We formulate model selection as an infinite-armed bandit problem. The models are arms, and picking an arm corresponds to a partial training of the model (resource allocation). The reward is the accuracy of the selected model after its partial training. In this best arm identification problem, regret is the gap between the expected accuracy of the optimal model and that of the model finally chosen. We introduce the algorithm Mutant-UCB, a generalization of UCB-E (see Audibert et al. 2010) to the stochastic infinite-armed bandit problem that incorporates operators from evolutionary algorithms. Tests were carried out on the optimization of deep neural networks (architectures and hyperparameters). The results, which outperform the state-of-the-art on three image classification datasets, demonstrate the relevance of our approach.

**Keywords.** Infinite-armed bandits, Model selection, Neural architecture optimisation, Hyperparameter optimisation, Evolutionnary algorithm, Image classification, AutoML, Online Learning

# 1 Introduction

Les performances des modèles de *machine learning* dépendent d'un grand nombre de paramètres qui ne peuvent pas être optimisés durant l'entraînement. Tester à la main toutes les combinaisons de ces paramètres est impossible. C'est pourquoi, les techniques permettant de sélectionner automatiquement ces paramètres, regroupées sous le nom d'AutoML (pour *Automated Machine Learning*), ont gagné en popularité (voir Hutter et al. 2019 pour un livre sur le sujet). Nous abordons l'optimisation de ces paramètres, aussi appelée sélection de modèles, d'une manière très générique. Nous ne nous restreignons à aucun type de paramètres à optimiser, qui peuvent être le modèle en lui-même, l'architecture d'un réseau de neurones ou les hyper-paramètres d'une forêt aléatoire par exemple. Notre but est de trouver le meilleur modèle pour une tâche donnée, sans faire d'hypothèse sur la tâche, le type de modèle ou la fonction de récompense à maximiser. Nous supposons uniquement que nous avons accès à un nombre infini de modèles potentiels. Pour réaliser notre optimisation, nous nous fixons un budget de $T$ ressources à utiliser pour entraîner nos modèles. Ces ressources sont allouées sous la forme d'entraînements partiels. Le modèle final est choisi en s'appuyant sur un compromis entre une bonne exploration (entraîner un grand nombre de modèles) et une bonne exploitation (allouer un budget de ressources conséquent aux modèles prometteurs). Nous nous plaçons dans le cadre des bandits avec une infinité de bras pour gérer ce compromis exploration-exploitation (voir Lattimore and Szepesvári 2020 pour une revue complète).

Dans ce papier, nous considérons ainsi la sélection de modèles comme l'identification du meilleur bras dans le cadre bandit avec un nombre infini de bras. Nous proposons de résoudre ce problème en considérant les algorithmes de type UCB (*Upper Confidence Bounds*, voir Auer et al. 2002). À partir de ces approches, nous avons créé un nouvel algorithme de sélection de modèles, appelé Mutant-UCB, qui utilise un opérateur de mutation venant des algorithmes évolutionnaires. Cet opérateur génère un nouveau modèle appartenant au voisinage du bras sélectionné par l'algorithme de bandits. Mutant-UCB ne fait aucune hypothèse sur la manière d'encoder les modèles à sélectionner (aussi appelé espace de recherche) ou sur la fonction de récompense à maximiser. Il est ainsi applicable à un très grand nombre de configurations. La combinaison d'un algorithme de type UCB avec une allocation adaptative du budget $T$ permet une bonne exploration de l'espace de recherche, tandis que l'opérateur de mutation oriente de manière efficace la recherche vers les solutions prometteuses. Nos résultats sur l'optimisation de réseaux de neurones démontrent l'intérêt de notre approche. La Section 2 présente le cadre bandit pour la sélection de modèles et nous positionne vis-à-vis de l'état de l'art. Dans la Section 3, nous détaillons notre algorithme Mutant-UCB. La Section 4 est dédiée à nos expérimentations sur l'optimisation de réseaux de neurones. Nous validons les performances de Mutant-UCB sur trois jeux de données de classification d'images disponibles en *open-source*. Enfin, la Section 5 discute des avantages de Mutant-UCB par rapport à l'état de l'art et ouvre de nouvelles perspectives de recherche.

# 2 Modélisation d'un problème de sélection de modèles : une approche bandit

## 2.1 État de l'art

Les premières stratégies pour la sélection de modèles sont la recherche en grille ou la recherche aléatoire (*Grid Search* et *Random Search*). Des méthodes plus sophistiquées abordent la sélection de modèles comme un problème d'apprentissage séquentiel. Deux approches se distinguent : la **sélection de configuration** (*configuration selection*) pour laquelle de nouveaux modèles "proches" des modèles prometteurs sont choisis séquentiellement, et l'**évaluation de configuration** (*configuration evaluation*, voir Li et al. 2018) qui alloue plus de ressources (temps d'entraînement) aux modèles prometteurs. La première approche suggère l'existence d'une distance entre les modèles (possiblement dans un espace latent sous-jacent), de sorte que deux modèles proches l'un de l'autre auront des performances similaires. La seconde approche, elle, n'introduit aucune notion de proximité entre les modèles, et, de fait, ne repose sur aucune hypothèse concernant leur performance.

**Algorithmes évolutionnaires.** Parmi les méthodes de sélection de modèles, les algorithmes évolutionnaires sont populaires depuis de nombreuses années (voir par exemple Young et al., 2015). Ils partent d'un ensemble de modèles initiaux, et les font évoluer vers des modèles performants en utilisant des opérateurs unitaires comme la mutation (peu de changement dans la configuration), ou impliquant plusieurs modèles comme le *crossover* (Strumberger et al., 2019). L'un des inconvénients des algorithmes évolutionnaires est le grand nombre de paramètres impliqués, tels que la taille de la population, la fonction de sélection ou le taux d'élitisme. Le choix des valeurs appropriées pour ces paramètres peut s'avérer complexe.

**Optimisation bayésienne.** L'optimisation bayésienne s'est récemment imposée comme une approche plus efficace que les méthodes évolutionnaires pour l'AutoML (voir, entre autres, Zoph and Le 2016). Cette technique d'optimisation séquentielle, utilisée pour la minimisation de fonctions boîtes noires repose sur deux composantes principales, un modèle de substitution (*surrogate model*) qui permet d'approximer la fonction objectif inconnue et une fonction d'acquisition qui permet de choisir le prochain élément de l'espace de recherche à évaluer. Les fonctions d'acquisition nécessitent un espace de recherche numérique voir continu (Garrido-Merchán and Hernández-Lobato, 2020). C'est pourquoi nous n'utilisons pas d'optimisation bayésienne dans nos expériences, car nous souhaitons un cadre plus générique.

**Algorithmes de bandits.** Toujours dans le domaine de l'optimisation bayésienne, des extensions de l'algorithme UCB classique ont été utilisés pour l'optimisation et la sélection de modèles tel que GP-UCB et KernelUCB (voir respectivement Srinivas et al. 2009 et Valko et al. 2013). Les méthodes d'évaluation de configurations ont également été étudiées dans un cadre de bandit à bras multiples ou infinis (*infinite* ou *multi-armed bandit*). À chaque itération de l'algorithme, un nouveau bras/modèle peut être tiré dans un espace de recherche infini contenant les modèles puis être ajouté à l'ensemble des modèles déjà (plus ou moins) entraînés. L'algorithme de *Sequential* (ou *Successive*) *Halving*, répartit le budget donné entre un certain

nombre (supposé optimal) de tours d'élimination, et tire de manière uniforme au sein d'un tour les bras à évaluer. Il a été étendu par Li et al. 2018, qui propose Hyperband, une extension plus robuste utilisée pour l'optimisation d'hyperparamètres de réseaux de neurones. Enfin, des méthodes hybrides combinant la sélection et l'évaluation de configurations adaptatives ont également été proposées par Kandasamy et al. 2016 qui étend GP-UCB pour permettre l'apprentissage séquentiel des modèles et l'allocation de ressources.

### 2.2 Contributions

Nous proposons un nouvel algorithme de sélection de modèles, Mutant-UCB, qui généralise l'algorithme UCB-E (Audibert et al., 2010) au problème de bandits avec une infinité de bras et incorpore un opérateur de mutation issu des algorithmes évolutionnaires. Notre algorithme combine les approches d'évaluation et de sélection des configurations : il repose sur une stratégie séquentielle et choisit un modèle grâce à un critère basé sur l'UCB. Ensuite, soit il continue l'entraînement du modèle sélectionné (allocation de ressources), soit il crée un nouveau modèle avec la mutation et l'entraîne. L'intuition est que la performance du "modèle mutant" sera proche de celle du modèle original. Bien que des approches bandit aient été utilisées pour l'opérateur de "sélection" d'un algorithme évolutionnaire (voir Li et al., 2013), c'est à notre connaissance la première fois qu'une mutation est incorporée dans un algorithme de bandit. Nous comparons les performances de Mutant-UCB avec une recherche aléatoire, un algorithme évolutionnaire et l'algorithme Hyperband (voir Li et al., 2018) sur trois jeux de données open source collectés pour la classification d'images. Pour une comparaison équitable, Mutant-UCB et l'algorithme évolutionnaire partagent le même opérateur de mutation.

### 2.3 Modélisation

À chaque itération $t \leq T$ de notre algorithme, un nouveau bras $k$ est choisi dans l'espace de recherche ou créé par mutation. Il est entraîné sur un jeu d'entraînement $\mathcal{D}_{\text{train}}$ puis évalué sur un jeu de validation $\mathcal{D}_{\text{valid}}$ à l'aide d'une fonction de précision (*accuracy*) acc qui correspond à la récompense $a_t$ que nous cherchons à maximiser. Dans ce qui suit, nous appelons $f_k$ un modèle non-entraîné associé au bras $k$, et $f_k^{N_k}$ lorsqu'il a été entraîné $N_k$ fois.

## 3 Mutant-UCB

L'Algorithme 1 détaille le pseudo-code de Mutant-UCB. Il échantillonne dans l'espace de recherche $K$ modèles initiaux $f_1, \ldots, f_K$, avec $K$ à fixer. Les $K$ modèles sont sous-entraînés une première fois donnant les précisions $a_k = \text{acc}(f_k^1, \mathcal{D}_{valid})$. Lors des itérations suivantes $t \leq T$, un même bras $k$ pourra être entraîné plusieurs fois, nous permettant de définir sa précision moyenne empirique $\hat{\mu}_k$ :

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{s=1}^{t-1} a_t \mathbf{1}_{I_s=k} \quad \text{avec} \quad N_k = \sum_{s=1}^{t-1} \mathbf{1}_{I_s=k} \,. \tag{1}$$

L'entier $N_k$ représente le nombre de fois où le bras $k$ a été choisi avant le tour $t$. La moyenne $\mu_k$ est ensuite utilisée par Mutant-UCB pour choisir de manière optimiste le bras suivant à évaluer, en résolvant l'équation :

$$I_t \in \underset{k\in\{1,\dots K\}}{\operatorname{argmax}} \left\{ \hat{\mu}_k + \sqrt{\frac{E}{N_k}} \right\}. \tag{2}$$

Avec $E$ un paramètre d'exploration à fixer. Nous introduisons à présent l'entier $\overline{N}_k$ qui compte le nombre de fois où le modèle associé au bras $k$ a été entraîné. Nous posons $N$ la valeur maximale que peuvent prendre les $N_k$. Une fois le bras $I_t$ choisi, en prenant $p_t = 1 - \overline{N}_{I_t}/N$, Mutant-UCB :

- effectue un sous-entraînement sur $f_{I_t}^{\overline{N}_{I_t}}$ avec une probabilité $p_t$ ou
- utilise l'opérateur de mutation sur $f_{I_t}^{\overline{N}_{I_t}}$ avec une probabilité $1 - p_t$.

La mutation est effectuée sur le modèle entraîné $f_{I_t}^{\overline{N}_{I_t}}$ et non uniquement sur $f_{I_t}$ afin d'inclure le cas où certains paramètres du modèle optimisés pendant l'entraînement (les poids d'un réseau de neurones pax exemple) seraient transmis à son mutant. Nous détaillons l'opération de mutation pour les expériences dans la Section 4. La mutation créée un nouveau modèle qui est immédiatement entraîné et ajouté à la liste des modèles potentiels à conserver à la fin de l'algorithme. Ainsi, le nombre de modèles $K$ augmente de 1 à chaque fois qu'un modèle mutant est créé. La probabilité $p_t$, elle, diminue au fur et à mesure que le modèle choisi a été entraîné. Il devient ainsi plus susceptible de muter. La probabilité limite le nombre de sous-entraînements d'un même modèle à $N$. L'idée sous-jacente est que multiplier les sous-entraînements sera inutile voir nuisible dans le cas des réseaux de neurones qui peuvent sur-apprendre. Si l'algorithme sélectionne ce modèle déjà bien entraîné, c'est parce qu'il est probablement performant et il faut donc chercher dans ses mutants pour trouver mieux. Notons que la probabilité $p_t$ est linéaire en $N_{I_t}$ ; ce choix est arbitraire et nous aurions très bien pu choisir un autre type de relation, par exemple, $p_t = 1 - \exp(N_{I_t} - N)$. Finalement, l'algorithme se termine par une dernière phase durant laquelle le meilleur modèle trouvé durant les $T$ itérations, $\widehat{I}_T$, est sélectionné et son entraînement est finalisé par $N - \overline{N}_{\widehat{I}_T}$ sous-entraînements supplémentaires.

*Remarque* 3.1. L'utilisation de l'opérateur de mutation repose sur l'intuition que la distribution des performances des mutants d'un certain modèle $j$ est différente de la distribution générale des performances dans notre espace de recherche, et une hypothèse sous-jacente pourrait être par exemple :

$$\mathbb{E}\Big[\, \mu_k \,\Big|\, f_k \text{ est un mutant de } f_j^{\overline{N}_j} \,\Big] = \mu_j \,.$$

# 4 Expériences

Dans cette section, nous évaluons les performances de Mutant-UCB, sur l'optimisation des réseaux de neurones. Afin de mettre en évidence les avantages de notre méthode, nous nous plaçons dans un cas où nous ne faisons aucune hypothèse sur la régularité de la récompense *acc* et où nous ne considérons aucune distance entre les éléments $f_k$ de notre espace de recherche.

**Algorithm 1** Mutant-UCB

**Paramètres** :
Budget $T$, paramètre d'exploration $E$, nombre de modèles $K$, nombre de sous-entraînements maximal pour un modèle $N$
**Initialisation**
Choisir aléatoirement $K$ modèles non-entraînés $f_1, \ldots, f_K$
**Pour** $k = 1, 2, \ldots, K$
Faire un premier sous-entraînement pour $f_k$, qui devient $f_k^1$ et récupérer $a_k = \mathrm{acc}\big(f_k^1, \mathcal{D}_{\text{valid}}\big)$
Définir $N_k = \overline{N}_k = 1$, $\hat{\mu}_k = a_k$
**Pour** $t = K+1, K+2, \ldots, (T-N+1)$
Prendre $I_t \in \mathrm{argmax}_{k\in\{1,\ldots K\}} \left\{\hat{\mu}_k + \sqrt{\frac{E}{\overline{N}_k}}\right\}$ et tirer $X_t \sim \mathcal{B}(p_t)$ avec $p_t = 1 - \overline{N}_{I_t}/N$
**Si** $X_t = 1$ :
Faire un sous-entraînement : $f_{I_t}^{\overline{N}_{I_t}}$ devient $f_{I_t}^{\overline{N}_{I_t}+1}$ et récupérer $a_t = \mathrm{acc}\big(f_{I_t}^{\overline{N}_{I_t}+1}, \mathcal{D}_{\text{valid}}\big)$
Mettre à jour $\hat{\mu}_{I_t} = \frac{1}{\overline{N}_{I_t}+1}\big(a_t + \overline{N}_{I_t}\hat{\mu}_{I_t}\big)$, $N_{I_t} = N_{I_t} + 1$ et $\overline{N}_{I_t} = \overline{N}_{I_t} + 1$
**Sinon** :
Mettre à jour $K = K + 1$ et créer $f_K$ depuis $f_{I_t}^{\overline{N}_{I_t}}$
Faire un premier sous-entraînement pour $f_K$ qui devient $f_K^1$ récupérer $a_t = \mathrm{acc}\big(f_K^1, \mathcal{D}_{\text{valid}}\big)$
Définir $N_K = \overline{N}_K = 1$, $\hat{\mu}_K = a_t$ et mettre à jour $N_{I_t} = N_{I_t} + 1$
**Finalisation**
Sélectionner le meilleur modèle $\widehat{I}_T \in \mathrm{argmax}_{k\in\{1,\ldots K\}} \hat{\mu}_k$ et effectuer $N - \overline{N}_{\widehat{I}_T}$ sous-entraînements
**Sortie** : $f_{\widehat{I}_T}^N$

Nous comparons donc nos méthodes à trois algorithmes applicables dans ce cas : une recherche aléatoire, l'algorithme Hyperband et un algorithme évolutionnaire. Cette optimisation des réseaux de neurones est appliquée à trois jeux de données de classification d'images.

## 4.1 Présentation des expériences

**Jeux de données.** Nous avons réalisé nos expériences à l'aide de trois jeux de données de classification d'images, utilisés par Li et al. 2018 pour présenter l'algorithme Hyperband : CIFAR-10 (Krizhevsky et al., 2009), Street View House Numbers (SVHN, voir Netzer et al., 2011) et une version d'MNIST tournées et avec des images d'arrière-plan, appelée MRBI (Larochelle et al., 2007). CIFAR-10 et SVHN contiennent des images RVB de $32 \times 32$, des images $28 \times 28$ en niveaux de gris pour MRBI. Les labels sont convertis en entiers entre 0 et 9. Chaque jeu de données est divisé en trois : le jeu d'entraînement, de validation et de test. Le jeu d'entraînement est utilisé pour optimiser les poids du modèle (c'est-à-dire pour effectuer les sous-entraînements) et celui de validation pour obtenir les précisions. Enfin, la précision des modèles finaux de chacun des algorithmes est calculée sur le jeu de test. CIFAR-10 comporte 35k images dans le jeu d'entraînement, 15k dans celui de validation et 10k dans celui de test, SVHN 51k, 22k et 26k et MRBI 10k, 2k et 50k. Pour tous les jeux de données, nous avons normalisé les images de manière à ce que les images aient une moyenne de zéro et un écart-type de un.

**L'espace de recherche** Nous avons utilisé le *framework DRAGON* développé par Keisler et al. 2023 pour représenter nos réseaux de neurones. L'article contient une explication et

un tutoriel pour utiliser le package associé. Dans ce *framework*, les réseaux de neurones sont représentés par des graphes acycliques dirigés (DAGs), où les nœuds représentent les couches (par exemple récurrentes, *feed-forward*, convolutives) et les arêtes représentent les connexions entre elles. La tâche sur laquelle nous voulons tester nos algorithmes est la classification d'images. Pour ce faire, nous définissons un espace de recherche (l'ensemble des modèles possibles $f_k$) dédié à cette tâche. Tout modèle échantillonné $f_k$ est ainsi composé de DAGs. Le premier traite des données 2D et peut être constitué de convolutions 2D, de *pooling* 2D, de normalisation et de *dropout*. Le second est constitué d'une couche de *flatten* suivie de couches MLP (*Multi-Layers Perceptrons*) et de normalisation. Une dernière couche MLP est ajoutée à la sortie du modèle pour convertir le vecteur latent dans le format de sortie souhaité. Le *framework* propose un certain nombre d'opérateurs, pour optimiser ces graphes, tels que des mutations et des *crossover*. Les opérateurs de mutation modifient l'architecture du réseau de neurones en ajoutant, supprimant ou modifiant les nœuds et les connexions dans le graphe. Ils peuvent également être appliqués à l'intérieur des nœuds, sur les hyperparamètres du réseau de neurones (par exemple, la taille du noyau de la couche de convolution ou une fonction d'activation). Le *crossover* consiste à échanger les parties de deux graphes.

**Sous-entraînements.** Nous avons entraîné nos réseaux de neurones à l'aide d'un taux d'apprentissage cyclique comme suggéré par Huang et al. 2017. Quand le taux d'apprentissage est faible, le réseau de neurones atteint un minimum local et lorsqu'il ré-augmente, le modèle en sort. Dans nos expériences, un sous-entraînement correspond à l'une de ces boucles, avec un taux d'apprentissage passant de son maximum à son minimum. Nous notons $N$ le nombre maximum de sous-entraînements par élément $f_k$ de notre espace de recherche.

***Baselines.*** La recherche aléatoire (RS), l'algorithme évolutionnaire (EA), Hyperband et Mutant-UCB ont tous été implémentés de manière à pouvoir être utilisés avec DRAGON. Ils utilisent les mêmes fonctions de d'entraînement et de validation pour évaluer les performances des réseaux de neurones et ont le même nombre de ressources $T$. Pour la recherche aléatoire, nous sélectionnons au hasard $K_{\mathrm{RF}} = T/N$ réseaux de neurones. Pour chacun d'entre eux, nous effectuons $N$ sous-entraînements. Pour l'algorithme évolutionnaire, nous avons implémenté une version asynchrone. Par rapport à l'algorithme standard, la version asynchrone est plus efficace dans un environnement HPC (*High Performance Computing*) car elle crée deux nouveaux individus dès qu'un processus est disponible, et n'attend pas que toute la population soit évaluée (Liu et al., 2018). Nous définissons une population initiale de taille $K_{\mathrm{EA}}$, de réseaux de neurones initiaux et nous les entraînons $N$ fois. Ensuite, nous faisons évoluer la population à l'aide des opérateurs de mutation et de *crossover* de DRAGON et générons au total $T/N - K_{\mathrm{EA}}$ descendants, tous entraînés $N$ fois. Si un descendant généré est meilleur que le plus mauvais modèle présent dans la population, il le remplace. Pour Hyperband, nous avons exécuté l'algorithme avec ses paramètres $R$ et $\eta$ de sorte que le nombre total de sous-entraînements soit $T$ et que chaque modèle ne puisse être entraîné plus de $N$ fois (voir Li et al., 2018 pour plus de détails). L'algorithme Mutant-UCB démarre avec une population initiale de $K_{\textsc{Mutant}}$ et utilise un budget de $T$. Pour une comparaison équitable, EA et Mutant-UCB utilisent les mêmes opérateurs de mutation. Nous fixons $K_{\mathrm{EA}} << K_{\textsc{Mutant}} \lesssim K_{\mathrm{RF}}$. En effet, chaque modèle est entièrement entraîné par l'algorithme évolutionnaire, $K_{\mathrm{EA}}$ doit être plus faible que $T/N$ pour permettre la génération d'un nombre suffisant de descendants. Avec

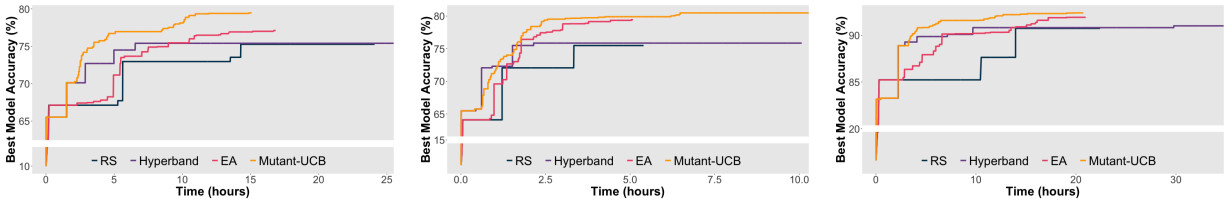
l'allocation de ressources, Mutant-UCB pourra générer et évaluer plus que $T/N$ modèle (c'est aussi le cas pour HyperBand), d'où $K_{\text{MUTANT}} >> K_{\text{EA}}$.

## 4.2 Résultats

TABLE 1 – Nombre de modèles testés et précision (en %) du meilleur modèle pour la recherche aléatoire (RS), l'algorithme évolutionnaire (EA) et Mutant-UCB sur CIFAR-10, MRBI et SVHN.

| Jeu de données | CIFAR-10 | MRBI | SVHN |
|---|---|---|---|
| RS | 1 000 · 75.3 | 1 000 · 75.5 | 1 000 · 90.7 |
| EA | 1 000 · 77.1 | 1 000 · 79.5 | 1 000 · 91.9 |
| Hyperband | 2 400 · 75.4 | 2 400 · 75.9 | 2 400 · 91.0 |
| Mutant-UCB | **3 399 · 79.5** | **3 463 · 80.5** | **3 471 · 92.4** |

Pour nos expériences, nous fixons $T = 10\,000$, $N = 10$ et $E = 0,05$ pour Mutant-UCB. Chaque sous-entraînement contient 10 époques, ce qui donne un maximum de 100 époques d'entraînement, et le taux d'apprentissage est fixé à 0,01. Chaque expérience est réalisée dans un environnement HPC avec de 20 GPUs NVIDIA V100. La Table 1 détaille les précisions (*accuracies*) maximales et le nombre de modèles testés pour chaque algorithme. Mutant-UCB bat les autres algorithmes pour tous les jeux de données. L'utilisation de la mutation semble être le facteur principal de cette performance puisque l'algorithme évolutionnaire devance largement Hyperband et la recherche aléatoire. Cependant, l'allocation de ressources joue également un rôle car Hyperband est meilleur que la recherche aléatoire et mutant meilleur que l'algorithme évolutionnaire, ce algorithmes convergent plus rapidement, comme on peut le constater sur la Figure 1. Les temps de calcul pour effectuer les $T$ itérations varient beaucoup entre les algorithmes et les jeux de données, d'abord pour des raisons matérielles, mais aussi à cause de l'allocation de ressources. Tout au long de l'article, nous supposons implicitement que le budget d'un sous-entraînement en terme de stockage et de temps de calcul est indépendant du modèle, ce qui est inexact. Un entraînement plus long de modèles complexes peut prendre plus de temps et affecte la durée totale de l'expérience. Mutant-UCB, avec l'opérateur de mutation et l'allocation des ressources, a la convergence la plus rapide et donne les meilleures précisions.




(a) CIFAR-10. (b) MRBI. (c) SVHN.

FIGURE 1 – Précision (*accuracy*) du meilleur modèle en fonction du temps de calcul pour la recherche aléatoire (RS), l'algorithme évolutionnaire (EA) et Mutant-UCB sur CIFAR-10, MRBI et SVHN.

# 5 Conclusion

Mutant-UCB, un algorithme de sélection de modèles innovant, qui combine un algorithme de bandit de type UCB avec un opérateur de mutation originaire des algorithmes évolutionnaires. La plupart des approches de sélection de modèless, telles que l'optimisation bayésienne ou les algorithmes de bandits continus, considèrent un espace vectoriel normé pour représenter l'ensemble des configurations possibles. Ces approches supposent que la fonction de récompense est suffisamment régulière sur cet espace normé, signifiant que deux configurations proches dans cet espace sous-jacent ont des précisions proches. Mutant-UCB et les autres algorithmes de nos expériences ne nécessitent pas ce type d'hypothèse. En outre, grâce à l'allocation des ressources, Mutant-UCB présente un potentiel exploratoire élevé. Il peut évaluer plus de modèles avec un même budget que la recherche aléatoire ou les algorithmes évolutionnaires. Par exemple, sur le jeu de données MRBI, avec un budget $T = 10\ 000$, Mutant-UCB a évalué 3 500 modèles, alors que l'algorithme évolutionnaire et la recherche aléatoire n'en ont évalué que 1000. L'utilisation de la mutation, qui peut être vue comme un concept de proximité sans nécessiter d'espace normé, renforce l'exploitation de solutions prometteuses et nous permet d'atteindre des modèles beaucoup plus performants que ceux trouvés par Hyperband et la recherche aléatoire. Elle reste plus permissive que le *crossover* qui nécessite l'homogénéité entre les éléments de l'espace de recherche. Ainsi, Mutant-UCB pourrait être utilisé avec un espace de recherche combinant différents modèles de machine learning, tels que les réseaux de neurones, les forêts aléatoires ou le *boosting*, à condition de définir un opérateur de mutation pour chaque type de modèle. Enfin, notre algorithme peut être implémenté de manière efficace sur HPC car les modèles sont évalués de manière indépendante et asynchrone, contrairement à Hyperband et aux algorithmes évolutionnaires classiques, qui évaluent les populations de manière synchrone. En résumé, l'algorithme Mutant-UCB présente plusieurs avantages qui en font un algorithme attrayant, en plus de ses performances démontrées dans la section précédente. La flexibilité de cet algorithme signifie qu'il peut être appliqué à un large éventail de problèmes. Une extension naturelle de cet article serait d'appliquer Mutant-UCB à une variété de tâches, de modèles et d'espaces de recherche où les algorithmes à l'état de l'art seraient limités voir inutilisables.

# Références


Audibert, J.-Y., S. Bubeck, and R. Munos (2010). Best arm identification in multi-armed bandits. In *COLT*, pp. 41–53.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*, 235–256.

Garrido-Merchán, E. C. and D. Hernández-Lobato (2020). Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing 380*, 20–35.

Huang, G., Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger (2017). Snapshot ensembles : Train 1, get m for free. *arXiv preprint arXiv :1704.00109*.

Hutter, F., L. Kotthoff, and J. Vanschoren (2019). *Automated machine learning : methods, systems, challenges*. Springer Nature.

Kandasamy, K., G. Dasarathy, J. B. Oliva, J. Schneider, and B. Póczos (2016). Gaussian process bandit optimisation with multi-fidelity evaluations. *Advances in neural information processing systems 29*.

Keisler, J., E.-G. Talbi, S. Claudel, and G. Cabriel (2023). An algorithmic framework for the optimization of deep neural networks architectures and hyperparameters. *arXiv preprint arXiv :2303.12797*.

Krizhevsky, A., G. Hinton, et al. (2009). Learning multiple layers of features from tiny images.

Larochelle, H., D. Erhan, A. Courville, J. Bergstra, and Y. Bengio (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pp. 473–480.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms*. Cambridge University Press.

Li, K., A. Fialho, S. Kwong, and Q. Zhang (2013). Adaptive operator selection with bandits for a multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation 18*(1), 114–130.

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar (2018). Hyperband : A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research 18*(185), 1–52.

Liu, H., K. Simonyan, O. Vinyals, C. Fernando, and K. Kavukcuoglu (2018). Hierarchical representations for efficient architecture search.

Netzer, Y., T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng (2011). Reading digits in natural images with unsupervised feature learning.

Srinivas, N., A. Krause, S. M. Kakade, and M. Seeger (2009). Gaussian process optimization in the bandit setting : No regret and experimental design. *arXiv preprint arXiv :0912.3995*.

Strumberger, I., E. Tuba, N. Bacanin, R. Jovanovic, and M. Tuba (2019). Convolutional neural network architecture design by the tree growth algorithm framework. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE.

Valko, M., N. Korda, R. Munos, I. Flaounas, and N. Cristianini (2013). Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv :1309.6869*.

Young, S. R., D. C. Rose, T. P. Karnowski, S.-H. Lim, and R. M. Patton (2015). Optimizing deep learning hyper-parameters through an evolutionary algorithm. In *Proceedings of the workshop on machine learning in high-performance computing environments*, pp. 1–5.

Zoph, B. and Q. Le (2016). Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*.

# Contrôle du taux de Fausses Découvertes pour les Knockoffs agrégés

Alexandre Blain [1] & Bertrand Thirion [2] & Olivier Grisel [3] & Pierre Neuvial [4]

[1] *INRIA, Université Paris-Saclay,* alexandre.blain@inria.fr
[2] *INRIA, CEA,* bertrand.thirion@inria.fr
[3] *INRIA,* olivier.grisel@inria.fr
[4] *Institut de Mathématiques de Toulouse, Université de Toulouse,*
pierre.neuvial@math.univ-toulouse.fr

**Résumé.** La sélection de variables contrôlée est une étape importante dans divers domaines scientifiques, tels que l'imagerie cérébrale ou la génomique. Dans ces contextes de données de haute dimension, considérer trop de variables conduit à des modèles médiocres et à des coûts élevés, d'où la nécessité de garanties statistiques sur les faux positifs. Les Knockoffs sont un outil statistique populaire pour la sélection conditionnelle de variables en haute dimension. Cependant, ils contrôlent l'espérance de la proportion de fausses découvertes (FDR) et non leur proportion réelle (FDP). Nous présentons une nouvelle méthode, KOPI, qui exploite la notion d'inférence post hoc pour contrôler les quantiles du FDP pour l'inférence basée sur les Knockoffs. La méthode proposée repose également sur un nouveau type d'agrégation pour contrer le caractère aléatoire indésirable associé à l'inférence Knockoff classique. Nous démontrons le contrôle du FDP et des gains de puissance substantiels par rapport aux méthodes basées sur les Knockoffs existantes dans divers contextes de simulation et obtenons de bons compromis sensibilité/spécificité sur des données d'imagerie cérébrale et génomique. Ce travail a fait l'objet d'un poster à la conférence NeurIPS 2023: https://arxiv.org/abs/2310.10373.

**Mots-clés.** Sélection de variable contrôlée, inférence Knockoffs, IRMf, génomique

**Abstract.** Controlled variable selection is an important analytical step in various scientific fields, such as brain imaging or genomics. In these high-dimensional data settings, considering too many variables leads to poor models and high costs, hence the need for statistical guarantees on false positives. Knockoffs are a popular statistical tool for conditional variable selection in high dimension. However, they control for the expected proportion of false discoveries (FDR) and not their actual proportion (FDP). We present a new method, KOPI, that controls the proportion of false discoveries for Knockoff-based inference. The proposed method also relies on a new type of aggregation to address the undesirable randomness associated with classical Knockoff inference. We demonstrate FDP control and substantial power gains over existing Knockoff-based methods in various simulation settings and achieve good sensitivity/specificity tradeoffs on brain imaging and genomic data. This work was pubished at NeurIPS 2023: https://arxiv.org/abs/2310.10373.

**Keywords.** Controlled variable selection, Knockoffs inference, fMRI, genomics

# 1 Inférence Knockoffs

**Notation.** Nous désignons les vecteurs par des lettres minuscules en gras. Un vecteur $\mathbf{x} = \{x_1, \ldots, x_p\}$ duquel nous avons retiré la $j^{me}$ coordonnée est désigné par $\mathbf{x}_{-j}$, c'est-à-dire $\mathbf{x} \setminus \{x_j\}$. L'indépendance entre deux vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$ est notée par $\mathbf{x} \perp \mathbf{y}$. Pour deux vecteurs $\mathbf{x}$ et $\tilde{\mathbf{x}}$ et un sous-ensemble $S$ d'indices, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(S)}$ désigne le vecteur obtenu à partir de $(\mathbf{x}, \tilde{\mathbf{x}})$ en échangeant les entrées $x_j$ et $\tilde{x}_j$ pour chaque $j \in S$. Les matrices sont notées par des lettres majuscules en gras, la seule exception étant le vecteur des statistiques Knockoff que nous notons par $\mathbf{W}$ comme dans [1, 6]. Pour tout ensemble $S$, $|S|$ désigne le cardinal de $S$. Pour un vecteur $\mathbf{z} = (z_j)_{1 \leq j \leq p}$ et $S \subset [\![p]\!]$, nous notons par $z_{(j:S)}$ (ou $z_{(j)}$ lorsqu'il n'y a pas d'ambiguïté) la $j^{me}$ plus petite valeur dans le sous-vecteur $(\mathbf{z}_s)_{s \in S}$. Pour un entier $k$, $[\![k]\!]$ désigne l'ensemble $\{1, \ldots, k\}$. L'égalité en distribution est notée par $\overset{d}{=}$.

**Le problème de sélection de variable conditionnelle.** Les données d'entrée sont notées par $\mathbf{X} \in \mathbb{R}^{n \times p}$, où $n$ est le nombre d'observations et $p$ le nombre de variables. La variable d'intérêt est noté par $\mathbf{y} \in \mathbb{R}^n$. L'objectif est de sélectionner les variables qui sont associées à la variable d'intérêt *conditionnellement à toutes les autres*. Formellement, nous testons simultanément pour tout $j \in [\![p]\!]$ :

$$H_{0,j} : y \perp x_j | \mathbf{x}_{-j} \quad \text{contre} \quad H_{1,j} : y \not\perp x_j | \mathbf{x}_{-j}.$$

La sortie d'une méthode de sélection de variables est un ensemble de rejet $\hat{S} \subset [\![p]\!]$ qui estime le support inconnu vrai $\mathcal{H}_1 = \{j : y \not\perp x_j | \mathbf{x}_{-j}\}$. Son complément est l'ensemble des vraies hypothèses nulles $\mathcal{H}_0 = \{j : y \perp x_j | \mathbf{x}_{-j}\}$. On note $p_0 = |\mathcal{H}_0|$. Pour assurer une inférence fiable, notre objectif est de fournir une garantie statistique sur la proportion de fausses découvertes dans $\hat{S}$. La Proportion de Fausses Découvertes (FDP) et le Taux de Fausse Découvertes (FDR) [2] sont définis comme suit :

$$\mathrm{FDP}(\hat{S}) = \frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}, \quad \mathrm{FDR}(\hat{S}) = \mathbb{E}[\mathrm{FDP}(\hat{S})] = \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\right].$$

Pour contrôler le FDP, nous utilisons la notion d'inférence post hoc introduite par [7]. Une borne supérieure post hoc de niveau $\alpha$ pour le FDP est une fonction $V$ qui vérifie :

$$\mathbb{P}\left(\forall S \subset [\![p]\!], \mathrm{FDP}(S) \leq V(S)/|S|\right) \geq 1 - \alpha\,.$$

**Knockoffs.** Le filtre Knockoff est une technique de sélection de variables introduite par [1] et affinée par [6] qui contrôle le FDR. Cette procédure repose sur la construction de copies bruitées des variables originales appelées variables Knockoff, qui sont conçues pour servir de contrôles pour la sélection de variables.

**Definition 1** (Model-X Knockoffs, 6)**.** Pour la famille de variables aléatoires $\mathbf{x} = (x_1, \ldots, x_p)$, les Knockoffs sont une nouvelle famille de variables aléatoires $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_p)$ satisfaisant :

1. pour tout $S \subset [\![p]\!]$, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(S)} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$

2. $\tilde{\mathbf{x}} \perp \mathbf{y}|\mathbf{x}$.

Une fois que nous disposons de telles variables, nous quantifions leur importance par rapport aux originales. Cela est fait en calculant les statistiques Knockoff $\mathbf{W} = (W_1, \ldots, W_p)$ qui sont définies comme suit.

**Definition 2** (Statistique Knockoff, 6)**.** Une statistique Knockoff $\mathbf{W} = (W_1, \ldots, W_p)$ est une mesure de l'importance des caractéristiques qui satisfait :

1. $\mathbf{W}$ dépend uniquement de $\mathbf{X}, \tilde{\mathbf{X}}$ et $\mathbf{y}$: $\mathbf{W} = g(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y})$.
2. Échanger la colonne $\mathbf{x}_j$ et sa colonne knockoff $\tilde{\mathbf{x}}_j$ inverse le signe de $W_j$:

$$W_j([\mathbf{X}, \tilde{\mathbf{X}}]_{swap(S)}, \mathbf{y}) = \begin{cases} W_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \text{ si } j \in S^c \\ -W_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \text{ si } j \in S. \end{cases}$$

La statistique Knockoff la plus couramment utilisée est la différence des coefficients Lasso (LCD) [18]. Cette statistique est obtenue en ajustant un estimateur Lasso [15] sur $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$, ce qui donne $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{2p}$. Ensuite, la statistique Knockoff peut être calculée en utilisant $\widehat{\boldsymbol{\beta}}$ :

$$\forall j \in [\![p]\!], \quad W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}|.$$

Ce coefficient résume l'importance de la $j^{me}$ variable originale par rapport à sa propre Knockoff : $W_j > 0$ indique que la variable originale est plus importante pour ajuster $y$ que la variable Knockoff, signifiant que la $j^{me}$ variable est probablement pertinente. Inversement, $W_j < 0$ indique que la $j^{me}$ variable est probablement non pertinente. Nous souhaitons donc sélectionner les variables correspondant aux $W_j$ grands et positifs. Formellement, l'ensemble de rejet $\hat{S}$ peut être écrit $\hat{S} = \{j : W_j > T_q\}$, où $T_q$ est choisi pour contrôler de manière prouvée le FDR au niveau $q$ [6].

**Schémas d'agrégation.** En raison de l'aléa inhérent au processus de génération des knockoffs, différentes variables peuvent être sélectionnées pour deux exécutions différentes de la méthode. Pour atténuer cela, l'agrégation de plusieurs tirages de Knockoffs est nécessaire. Ren and Barber [14] a introduit un schéma d'agrégation qui repose sur la définition des $e$-values Knockoffs.

$$e_j = \frac{p}{1 + |\{k : W_k \leq -T_q\}|} 1_{\{W_j \geq T_q\}}.$$

Ces $e$-values peuvent être moyennées sur $D$ tirages et la procédure e-BH [16] est effectuée pour la sélection de variables. Alternativement, [13] définit la $\pi$-statistique suivante, qui quantifie les preuves contre une variable :

$$\pi_j = \begin{cases} \frac{1+|\{k:W_k \leq -W_j\}|}{p} \text{ si} & W_j > 0 \\ 1 \quad \text{si} & W_j \leq 0. \end{cases} \tag{1}$$

Dans [13] les $\pi$-statistiques sont traitées comme des $p$-valeurs et agrégées en utilisant l'agrégation quantile [10]. Cependant, elles ne peuvent être considérées comme des $p$-valeurs que sous des hypothèses restrictives difficiles à vérifier. Dans la section suivante, ces statistiques sont utilisées comme un bloc de construction pour atteindre le contrôle du FDP. Le cadre KOPI ne nécessite pas que les $\pi$-statistiques soient des $p$-valeurs valides.

# 2 KOPI: Contrôle du taux de Fausses Découvertes pour les Knockoffs agrégés

La méthode que nous proposons vise à résoudre les deux problèmes principaux des méthodes Knockoffs existantes: *i)* le contrôle du FDR et non du FDP, qui ne sont pas des quantités équivalentes [11] et *ii)* le caractère aléatoire de l'inférence Knockoffs dû au processus de génération des Knockoffs.

## 2.1 Contrôle post hoc du FDP pour les $\pi$-statistiques

Pour obtenir un contrôle du FDP, nous nous appuyons sur le contrôle du Joint Error Rate (JER) tel qu'introduit dans [5]. Pour $k_{max} \in [\![p]\!]$, nous définissons une *famille de seuils* de taille $k_{max}$ comme un vecteur $\mathbf{t} = (t_j)_{j \in [\![k_{max}]\!]}$ tel que $0 \leq t_1 \leq \cdots \leq t_{k_{max}} \leq 1$.

**Definition 3** (Joint Error Rate, 5)**.** Soit $\pi_{(j:\mathcal{H}_0)}$ la $j^{me}$ plus petite valeur $\pi_j$ parmi toutes les hypothèses nulles. Le JER associé à $\mathbf{t} = (t_j)_{j \in [\![k_{max}]\!]}$ est :

$$\mathrm{JER}(\mathbf{t}) = \mathbb{P}\left(\exists j \in [\![k_{max} \wedge p_0]\!] : \pi_{(j:\mathcal{H}_0)} < t_j\right). \tag{2}$$

On dit que la famille de seuils $\mathbf{t}$ contrôle le JER au niveau $\alpha$ ssi $\mathrm{JER}(\mathbf{t}) \leq \alpha$.

Une borne supérieure de niveau $\alpha$ pour le FDP peut être déduite du contrôle du JER via le résultat suivant :

**Proposition 1** (Contrôle du FDP via contrôle du JER 5)**.** *Si* $\mathbf{t}$ *est une famille de seuils de taille* $k_{max}$ *qui contrôle le* JER *au niveau* $\alpha$*, alors,* $V^{\mathbf{t}}(S)/|S|$ *est une borne supérieure de niveau* $\alpha$ *pour le FDP, avec :*

$$V^{\mathbf{t}}(S) = \min_{1 \leq k \leq k_{max}} (k-1) + \sum_{i \in S} 1_{\{\pi_i > t_k\}}. \tag{3}$$

Dans la suite de cette section, nous montrons comment contrôler le JER dans le cas des $\pi$-statistiques.

## 2.2 Distribution conjointe des $\pi$-statistiques sous l'hypothèse nulle

Par la Définition 3, le $\mathrm{JER}(\mathbf{t})$ d'une famille de seuils donnée dépend uniquement de la distribution conjointe nulle des $\pi$-statistiques. Comme pour les résultats antérieurs de contrôle

du FDR [1] ou de contrôle du k-FWER [8], l'idée clé pour obtenir un contrôle du JER pour les $\pi$-statistiques est de prouver que la partie pertinente de cette distribution est en fait connue, grâce aux propriétés des statistiques knockoff. Nous utilisons la même notation que dans [8]. Soit $Z_j = |\{k \in [\![p]\!] : W_k \leq -W_j\}|$ et $\chi_j = sign(W_j)$, les $\pi$-statistiques $(\pi_j)_{j=[\![p]\!]}$ sont alors données par :

$$\pi_j = \frac{1+Z_j}{p} 1_{\{\chi_j=1\}} + 1_{\{\chi_j=-1\}}.$$

Pour un $\mathbf{W}$ donné, soit $\sigma(\mathbf{W})$ être une permutation de $[\![p]\!]$ qui trie $\mathbf{W}$ par module décroissant : $\sigma(\mathbf{W}) = (\sigma_1, \ldots, \sigma_p)$ tel que $|W_{\sigma_1}| \geq |W_{\sigma_2}| \cdots \geq |W_{\sigma_p}|$. Nous commençons par prouver que les statistiques $Z$ peuvent être exprimées comme une fonction du vecteur de statistiques $\chi$ :

**Lemma 1.** *Pour $j \in [\![p]\!]$ tel que $\chi_{\sigma_j} = 1$, $Z_{\sigma_j} = \sum_{k=1}^{j-1} 1_{\{\chi_{\sigma_k}=-1\}}$.*

*Preuve du Lemme 1.* Puisque $\chi_{\sigma_j} = 1$, nous avons :

$$\begin{aligned}
Z_{\sigma_j} &= |\left\{k \in [\![p]\!] : W_{\sigma_k} \leq -W_{\sigma_j}\right\}| \\
&= |\left\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ et } W_{\sigma_k} \leq -W_{\sigma_j}\right\}| \\
&= |\left\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ et } |W_{\sigma_k}| \leq |W_{\sigma_j}|\right\}| \\
&= |\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ et } k \leq j\}| \\
&= \sum_{k=1}^{j-1} 1_{\{\chi_{\sigma_k}=-1\}}.
\end{aligned}$$

□

Le Lemme 1 implique que la distribution des statistiques d'ordre de $\pi|\sigma(\mathbf{W})$ est entièrement déterminée par celle de $\chi|\sigma(\mathbf{W})$. Pour formaliser cela, nous introduisons les statistiques $\pi^0$.

**Definition 4** (Statistiques $\pi^0$)**.** Soit $\chi^0 = (\chi_j^0)_{1\leq j\leq p}$ une collection de $p$ variables aléatoires de Rademacher i.i.d., c'est-à-dire, pour tout $j$, $\mathbb{P}(\chi_j^0 = 1) = \mathbb{P}(\chi_j^0 = -1) = 1/2$. Les statistiques $\pi^0$ associées sont définies pour $j \in [\![p]\!]$ par

$$\pi_j^0 = \frac{1+Z_j^0}{p} 1_{\{\chi_j^0=1\}} + 1_{\{\chi_j^0=-1\}}, \text{ où } Z_j^0 = \sum_{k=1}^{j-1} 1_{\{\chi_k^0=-1\}}. \tag{4}$$

**Theorem 1.** *Soit $\mathbf{t}$ une famille de seuils de taille $k_{max}$. Alors, pour $\pi^0 = (\pi_j^0)_{j\in[\![p]\!]}$ comme dans* (4)*,*

$$\mathrm{JER}(\mathbf{t}) \leq \mathrm{JER}^0(\mathbf{t}) := \mathbb{P}\left(\exists k \in [\![k_{max}]\!] : \pi_{(k)}^0 < t_k\right). \tag{5}$$

Le Théorème 1 – prouvé dans le papier [3] – est lié au Lemme 3.1 de Janson and Su [8] et au Lemme 3.1 de Li et al. [9], qui s'appuient sur la propriété de changement de signe des statistiques Knockoff sous l'hypothèse nulle [1]. L'intérêt du Théorème 1 est que la borne supérieure $\mathrm{JER}^0(\mathbf{t})$ dépend uniquement des statistiques $\pi^0$ et de la famille de seuils $\mathbf{t}$, et non des données originales. Par conséquent, elle peut être estimée avec une précision arbitraire pour tout $\mathbf{t}$ donné en utilisant une simulation Monte-Carlo, comme expliqué dans la section suivante et décrit dans dans le Supp. Mat du papier [3].

## 2.3 Contrôle du Joint Error Rate pour les $\pi$-statistiques par calibration

Pour approcher la borne supérieure du JER dérivée dans le Théorème 1, nous tirons $B$ échantillons Monte-Carlo en utilisant l'Algorithme 1 du papier [3]. On obtient ainsi un ensemble de $B$ vecteurs de statistiques $\pi^0$ notés par $\pi^0_b \in \mathbb{R}^p$ pour chaque $b \in [\![B]\!]$. Cela nous permet d'évaluer le JER empirique, qui estime la borne supérieure d'intérêt.

**Definition 5** (JER Empirique)**.** Pour $B$ vecteurs de statistiques $\pi^0$ et une famille de seuils $\mathbf{t}$, le JER empirique est défini comme :

$$\widehat{\mathrm{JER}}^0_B(\mathbf{t}) = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\exists k \in [\![k_{max}]\!] : \pi^0_{b(k)} < t_k\right\}, \tag{6}$$

où pour chaque $b \in [\![B]\!]$, $\pi^0_{b(1)} \leq \cdots \leq \pi^0_{b(p)}$.

Puisque $\widehat{\mathrm{JER}}^0_B(\mathbf{t})$ peut être rendu arbitrairement proche (en choisissant $B$ assez grand) de $\widehat{\mathrm{JER}}^0(\mathbf{t})$ pour toute famille de seuils $\mathbf{t}$ donnée, il reste à choisir $\mathbf{t}$ tel que $\widehat{\mathrm{JER}}^0(\mathbf{t}) \leq \alpha$ afin d'assurer un contrôle du JER. À cette fin, nous considérons un ensemble trié de familles de seuils candidates appelé un *template* :

**Definition 6** (Template [5])**.** Un template est une fonction non décroissante composant par composant $\mathbf{T} : [0,1] \mapsto \mathbb{R}^p$ qui associe un paramètre $\lambda \in [0,1]$ à une famille de seuils $\mathbf{T}(\lambda) \in \mathbb{R}^p$.

Cette définition peut naturellement s'étendre au cas de templates contenant un nombre fini de familles de seuils. Le template correspondant à $B'$ familles de seuils est alors noté par $(\mathbf{T}(b'/B'))_{b' \in [\![B']\!]}$.

Une fois un template spécifié, la procédure de *calibration* [5] peut être appli ; cela consiste à trouver la famille de seuils la moins conservatrice $\mathbf{t}$ du template, parmi celles qui contrôlent le JER empirique au niveau $\alpha$. Formellement, nous considérons la famille de seuils définie $\mathbf{t}^B_\alpha = \mathbf{T}(\lambda_B(\alpha))$, où

$$\lambda_B(\alpha) = \frac{1}{B'}\max\left\{b' \in [\![B']\!] \quad s.t. \quad \widehat{\mathrm{JER}}^0_B\left(\mathbf{T}\left(\frac{b'}{B'}\right)\right) \leq \alpha\right\}.$$

Comme observé par Blain et al. [4], une puissance optimale est atteinte lorsque les familles candidates correspondent à la forme de la distribution des statistiques nulles. Nous définissons un template basé sur la distribution des statistiques $\pi^0$ apparaissant dans le Théorème 1. En pratique, nous tirons $B'$ échantillons de cette distribution indépendamment des $B$ échantillons Monte Carlo pour éviter les biais de circularité. Puisqu'un template doit être non décroissant composant par composant, c'est-à-dire que l'ensemble des familles de seuils candidates doit être trié, nous extrayons des quantiles empiriques de ces $B'$ vecteurs triés. Cela produit un template $\mathbf{T}^0$ composé de $B'$ courbes candidates qui correspondent aux quantiles de la distribution des statistiques $\pi^0$. La courbe de quantile $\frac{b'}{B'}$ définit la famille de seuils $\mathbf{T}^0(b'/B')$. Nous obtenons le résultat suivant :

**Theorem 2** (Contrôle du JER pour les $\pi$-statistiques)**.** *Considérons la famille de seuils définie par* $\mathbf{t}_\alpha^B = \mathbf{T}^0(\lambda_B(\alpha))$*. Alors, lorsque* $B \to +\infty$*,*

$$\mathrm{JER}(\mathbf{t}_\alpha^B) \leq \alpha + O_P(1/\sqrt{B}).$$

Le nombre $B$ d'échantillons Monte-Carlo dans le Théorème 2 peut être choisi arbitrairement grand pour obtenir un contrôle du JER, conduisant à des bornes FDP valides via l'Équation 3. Ce résultat est prouvé dans le papier [3].

Cette approche s'étend naturellement au cas agrégé comme montré dans le papier [3].

# 3 Quelques résultats expérimentaux

**Méthodes considérées.** Dans notre mise en œuvre de KOPI, nous nous appuyons sur la moyenne harmonique [19] comme schéma d'agrégation $f$. De plus, nous fixons $k_{max} = \lfloor p/50 \rfloor$ suivant l'approche de [4]. Nous considérons également les schémas d'agrégation de Knockoffs de l'état de l'art: AKO (Aggregation of Multiple Knockoffs, 13) et l'agrégation basée sur les $e$-values [14]. En outre, nous considérons les Knockoffs standard ("Vanilla knockoffs)", c'est-à-dire [6] et le contrôle du FDP via le Closed Testing [9]. Dans les expériences sur des données simulées, nous générons des Knockoffs en supposant une distribution gaussienne pour $\mathbf{X}$, avec toutes les variables centrées. Pour les méthodes qui permettent l'agrégation, nous utilisons $D = 50$ tirages de Knockoffs.

## 3.1 Données simulées

**Cadre de simulation.** À chaque simulation, nous générons des données gaussiennes $\mathbf{X} \in \mathbb{R}^{n \times p}$ avec une matrice de corrélation de Toeplitz correspondant à un modèle auto-régressif d'ordre 1, de paramètre $\rho$, c'est-à-dire $\mathbf{\Sigma}_{i,j} = \rho^{|i-j|}$.

Ensuite, nous tirons le vrai support $\boldsymbol{\beta}^* \in \{0,1\}^p$. Le nombre de coefficients non nuls de $\boldsymbol{\beta}^*$ est contrôlé par le paramètre de parcimonie $s_p$, c'est-à-dire $s_p = \|\boldsymbol{\beta}^*\|_0/p$. La variable cible $\mathbf{y}$ est construite à l'aide d'un modèle linéaire :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon},$$

avec $\sigma$ contrôlant l'amplitude du bruit : $\sigma = \|\mathbf{X}\boldsymbol{\beta}^*\|_2/(\mathrm{SNR}\|\boldsymbol{\epsilon}\|_2)$, SNR étant le rapport signal à bruit. Nous choisissons le paramètre central $n = 500, p = 500, \rho = 0.5, s_p = 0.1, \mathrm{SNR} = 2$. Pour chaque paramètre, nous explorons une gamme de valeurs possibles pour comparer les méthodes dans divers paramètres.

Pour sélectionner les variables en utilisant les bornes supérieures du FDP, nous retenons l'ensemble le plus grand possible de variables $S$ tel que $V(S) \leq q|S|$. Pour chacune des $N$ simulations et chaque méthode, nous calculons le FDP empirique et la Proportion de Vrais Positifs (TPP) :

$$\widehat{FDP}(S) = \frac{|S \cap \mathcal{H}_0|}{|S|} \quad \text{et} \quad \widehat{TPP}(S) = \frac{|S \cap \mathcal{H}_1|}{|\mathcal{H}_1|}.$$

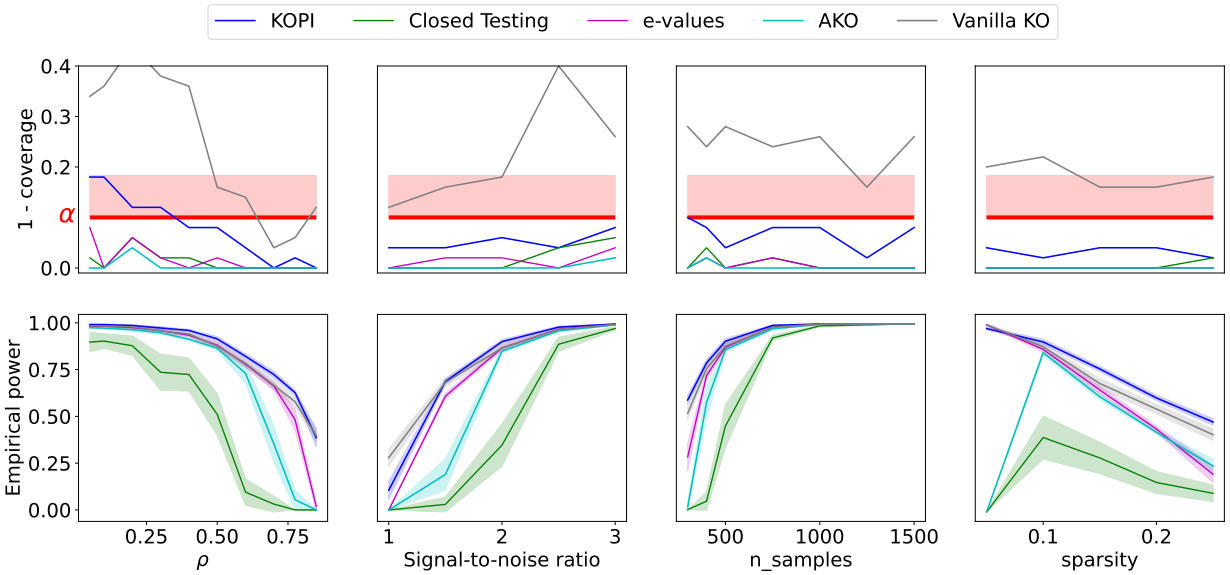



Figure 1: **Couverture de la borne du FDP au niveau $\alpha$ et puissance empirique pour 50 exécutions de simulation et cinq méthodes différentes :** Knockoffs standard, Knockoffs agrégés utilisant les e-values, Knockoffs agrégés utilisant l'agrégation quantile, KOPI et l'inférence Knockoff via le Closed Testing. Nous utilisons $D = 50$ tirages de Knockoffs et les paramètres de simulation suivants : $\alpha = 0.1, q = 0.1, p = 500$. Chaque colonne correspond à une expérience ou l'on fait varier le paramètre indiqué en abscisse, la première ligne affichant la couverture du FDP et la deuxième ligne la puissance. La ligne rouge et les bandes d'erreur associées représentent les limites acceptables (au niveau $\alpha$) pour la couverture de la borne du FDP. KOPI surpasse constamment toutes les autres méthodes tout en conservant le contrôle du FDP.

Si le FDP est contrôlé au niveau $\alpha$, $|\{k \in [\![N]\!] : \widehat{FDP}(S_k) > q\}| \sim \mathcal{B}(N, \alpha)$. Ensuite, nous pouvons calculer les bandes d'erreur sur le niveau $\alpha$ en utilisant $\mathrm{std}\left(\mathcal{B}(N, \alpha)/N\right) = \sqrt{\alpha(1-\alpha)/N}$. La deuxième ligne de la Fig. 1 représente la puissance empirique atteinte par chaque méthode, qui correspond à la moyenne des TPPs définis ci-dessus pour $N$ simulations, c'est-à-dire $\text{Puissance} = \sum_{k=1}^{N} \widehat{TPP}(S_k)/N$. La Fig. 1 montre que dans tous les paramètres différents, KOPI conserve le contrôle du FDP. Nous pouvons également voir que le contrôle du FDR n'implique pas le contrôle du FDP, car les Knockoffs standard sont systématiquement en dehors des intervalles de couverture de la borne du FDP. Cependant, les deux schémas d'agrégation existants (AKO et e-valeurs) qui garantissent formellement le contrôle du FDR sont généralement conservateurs et contrôlent empiriquement le FDP. Ceci est cohérent avec les conclusions de [14]. La procédure de Closed Testing de [9] contrôle le FDP comme annoncé mais souffre d'un manque de puissance.

Nous constatons ici que KOPI contrôle le FDP tout en offrant des gains de puissance par rapport aux méthodes d'agrégation de Knockoffs contrôlant le FDR. Pourtant, le contrôle du FDP est une garantie plus stricte que le contrôle du FDR, comme discuté précédemment. Ces gains sont particulièrement notables dans les cadres de simulations les plus difficiles, où la plupart des autres méthodes montrent une diminution claire de la puissance ou même un comportement catastrophique (c'est-à-dire une puissance nulle).

### 3.2 Application aux données d'IRMf

L'objectif de la cartographie du cerveau humain est d'associer des tâches cognitives à des régions cérébrales pertinentes. Ce problème est abordé à l'aide de l'Imagerie par résonance magnétique fonctionnelle (IRMf), qui consiste à enregistrer le niveau d'oxygénation du sang dans le cerveau à l'aide d'un scanner IRM.

L'importance de l'inférence conditionnelle pour ce problème a été soulignée dans [17]. Nous utilisons l'ensemble de données du Projet Connectome Humain (HCP900) qui contient des images cérébrales de jeunes adultes en bonne santé effectuant différentes tâches tout en étant dans un scanner IRM. Les détails sur cet ensemble de données et les résultats empiriques peuvent être trouvés dans l'Annexe D du papier [3].

Le contrôle du FDP et la puissance ne peuvent pas être évalués dans l'application précédente, puisque la vérité terrain n'est pas connue. Par conséquent, suivant l'approche de [12], nous avons effectué une expérience supplémentaire qui consiste à utiliser des données semi-simulées. Nous considérons un premier ensemble de données fMRI $(\mathbf{X}_1, \mathbf{y}_1)$ sur lequel nous effectuons une inférence en utilisant un estimateur Lasso ; cela produit $\boldsymbol{\beta}_1^* \in \mathbb{R}^p$ que nous utiliserons comme notre vérité terrain. Ensuite, nous considérons un ensemble de données fMRI différent $(\mathbf{X}_2, \mathbf{y}_2)$ pour la génération de données. L'intérêt d'utiliser un ensemble de données différent est d'éviter un biais de circularité entre la définition de la vérité terrain et la procédure d'inférence. Concrètement, nous écartons le vecteur de réponse original $\mathbf{y}_2$ pour cet ensemble de données et construisons une réponse simulée $\mathbf{y}_2^{sim}$ en utilisant un modèle linéaire, avec la même notation que précédemment (nous fixons $\sigma$ de sorte que SNR $= 4$) : $\mathbf{y}_2^{sim} = \mathbf{X}_2\boldsymbol{\beta}_1^* + \sigma\boldsymbol{\epsilon}$.

Ensuite, l'inférence knockoff est réalisée à partir des données $(\mathbf{X}_2, \mathbf{y}_2^{sim})$. Puisque nous considérons $\boldsymbol{\beta}_1^*$ comme la vérité terrain, le FDP et le TPP peuvent être calculés pour chaque méthode. Comme on peut le voir dans la Fig. 2, KOPI est la méthode la plus puissante parmi celles qui contrôlent le FDP.

## 4 Discussion

Nous avons proposé une nouvelle méthode qui contrôle du FDP sur les Knockoffs agrégés. Elle combine les avantages de l'agrégation, c'est-à-dire l'amélioration de la stabilité de l'inférence, en plus de fournir un contrôle probabiliste du FDP, plutôt que de contrôler seulement son espérance, le FDR.

Les résultats de simulation confirment que KOPI contrôle effectivement le FDP. De plus, bien que le contrôle du FDP soit une garantie plus stricte que le contrôle du FDR, KOPI offre en réalité des gains de puissance par rapport à l'état de l'art pour l'agrégation de Knockoffs.

Un package Python contenant le code pour KOPI est disponible à l'adresse: `https://github.com/alexblnn/KOPI`.

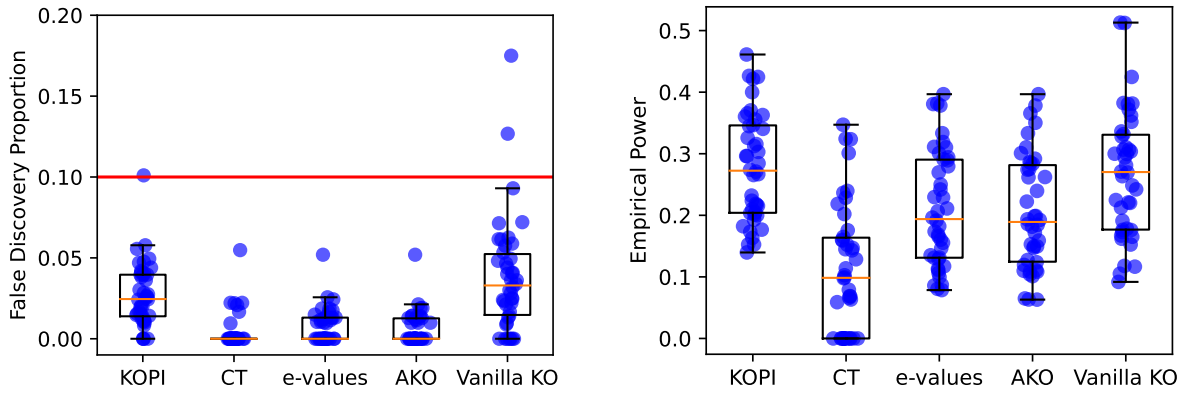



Figure 2: **FDP empirique et puissance sur des données semi-simulées pour 42 paires de contrastes.** Nous utilisons 7 contrastes HCP C0 : "Main motrice", C1 : "Pied motrice", C2 : "Jeu", C3 : "Relationnel", C4 : "Émotion", C5 : "Social", C6 : "Mémoire de travail". Nous considérons toutes les 42 paires d'entraînement/test possibles : le contraste d'entraînement est utilisé pour obtenir une vérité terrain, tandis que le contraste de test est utilisé pour générer la réponse. L'inférence est effectuée en utilisant les 5 méthodes considérées dans l'article et le FDP empirique est rapporté dans le diagramme à boîte de gauche, tandis que la puissance est rapportée dans le diagramme à boîte de droite. Remarquez (figure de droite) que KOPI offre une puissance supérieure par rapport à toutes les autres méthodes basées sur les Knockoffs tout en contrôlant le FDP (fig. de gauche).

# References

[1] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

[3] Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs. *NeurIPS 2023*.

[4] Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492.

[5] Blanchard, G., Neuvial, P., Roquain, E., et al. (2020). Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*, 48(3):1281–1303.

[6] Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

[7] Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597.

[8] Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1):960–975.

[9] Li, J., Maathuis, M. H., and Goeman, J. J. (2022). Simultaneous false discovery proportion bounds via knockoffs and closed testing. *arXiv preprint arXiv:2212.12822*.

[10] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

[11] Neuvial, P. (2020). *Contributions to statistical inference from genomic data*. Habilitation thesis, Université Toulouse III Paul Sabatier.

[12] Nguyen, B. T., Thirion, B., and Arlot, S. (2022). A Conditional Randomization Test for Sparse Logistic Regression in High-Dimension. In *NeurIPS 2022*, volume 35 of *Advances in Neural Information Processing Systems*, New Orleans, United States.

[13] Nguyen, T.-B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2020). Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR.

[14] Ren, Z. and Barber, R. F. (2022). Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*.

[15] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[16] Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852.

[17] Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59.

[18] Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*.

[19] Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.

# Rethinking multiple kernel learning under the lenses of stochastic variational inference

Davide Adamo [1,2] & Marco Corneli [1,2]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France.*

[2] *Université Côte d'Azur, Laboratoire CEPAM, Nice, France.*
*davide.adamo@univ-cotedazur.fr*

**Résumé.** Les méthodes de noyaux sont largement utilisées dans l'apprentissage automatique car elles constituent un outil puissant pour cartographier implicitement les données dans des espaces à haute dimension, permettant la découverte de modèles complexes qui pourraient être difficiles à capturer dans l'espace de caractéristiques d'origine. Bien que de nombreux problèmes de classification et de régression puissent être résolus avec succès à l'aide d'un seul noyau, il arrive que les ensembles de données du monde réel présentent des structures diverses et qu'il soit nécessaire d'utiliser plusieurs types de noyaux (un pour chaque notion de similarité que l'on souhaite prendre en compte). C'est là que l'apprentissage multi-noyaux (MKL) entre en jeu.
Cet article revisite la classification multi-noyaux en mettant l'accent sur la sélection des noyaux à la lumière des développements récents de l'inférence variationnelle stochastique (SVI). Dans le cadre de la régression logistique à noyaux, nous considérons des combinaisons linéaires semi-définies positives de noyaux et nous traitons les poids des noyaux comme des variables aléatoires. Des choix appropriés de distributions préalables font naturellement émerger une pénalité Lasso, tandis que la puissance du SVI nous permet d'estimer le modèle et les paramètres variationnels dans un contexte entièrement différentiable, ainsi que de construire des intervalles de confiance pour les poids des noyaux. Des exemples numériques illustrent notre approche.

**Mots-clés.** Apprentissage à noyaux multiples, Inférence variationnelle stochastique, sélection du modèle.

**Abstract.** Kernel methods are widely employed in machine learning as they are a powerful tool to implicitly map data into high-dimensional spaces, enabling the discovery of complex patterns that might be challenging to capture in the original feature space. Although many classification and regression problems can be successfully attacked with a single kernel, sometimes real-world datasets exhibit diverse structures and employing several kernel types (one for each notion of similarity that we aim to take into account) is necessary. This is where multi-kernel learning (MKL) comes into play.
This paper revisits multi-kernel classification with a specific focus on kernel(s) selection in the light of the recent developments in stochastic variational inference (SVI). In the framework of kernelized logistic regression, we consider positive semi-definite linear combitinations of kernels and we treat the kernel weights as random variables. Proper choices of prior distributions make naturally emerge a Lasso penalty, whereas the power of SVI allows us to

estimate the model and variational parameters in a fully differentiable context as well as to build confidence intervals for the kernel weights. Numerical examples highlight our approach.

**Keywords.** Multiple kernel learning, Stochastic variational inference, Model selection.

# 1 Introduction

Kernel methods provide a powerful framework for capturing complex relationships in data by implicitly mapping input features into higher-dimensional spaces, the reproducing kernel Hilbert spaces (RKHS). This transformation both enables linear models to operate effectively despite non-linear feature domains and boost their ability to model intricate patterns and structures. The key ingredient in kernel methods is the kernel (Vapnik, 1999) itself, a symmetric positive semi-definite function that determines the similarity between pairs of data points. Standard kernels are the linear, polynomial, and radial basis function (RBF) one, however an infinite number of kernels can be manufactured and choosing the "right" one(s) is crucial to boost the performance of a classification or regression model. Multiple Kernel Learning (MKL) (Sonnenburg et al., 2006; Gönen and Alpaydın, 2011) emerges as an attempt to both exploit the information coming from different kernels and (possibly) select the relevant kernels for a given machine learning task. This challenge is addressed by allowing the integration of different input kernels and finding an optimal linear or non-linear combination of such kernels.

In (Rakotomamonjy et al., 2008) was developed Simple MKL, using a sub-gradient descent (SD) approach to handle high-dimensional data. This method involves the strategic selection of a subset of pertinent kernels, leading to an improvement in computational efficiency while preserving the advantages associated with integrating multiple kernels. These methods tend to produce sparse kernel combinations ($l1$ penalty on kernel weights) although, being pure optimization any form of inference on the kernel weighs is prevented. In the Bayesian universe, (Girolami and Rogers, 2005) introduced a Bayesian MKL approach for binary classification using hierarchical models and (Damoulas and Girolami, 2008) extended it to a multiclass formulation. In those works, a convex sum of input kernels was enforced using a Dirichlet prior on the kernel weights. Given the high computational cost of training these algorithms, (Gönen, 2012) opted for a fully conjugate Bayesian formulation (BEMKL), involving Gaussian processes. A recent review of the existing techniques that bridge MKL and deep learning approaches can be found in (Wang et al., 2021).

In the light of some (relatively) recent developments in stochastic variational inference (Blei et al., 2017; Naesseth et al., 2017) this paper revisits MKL in the context of supervised classification with the aim to preserve and possibly improve the simplicity of the approach described in (Rakotomamonjy et al., 2008) for the kernel(s) selection, whereas exploiting the versatility of Bayesian approaches, in particular allowing one to perform posterior inference on the kernel weights.

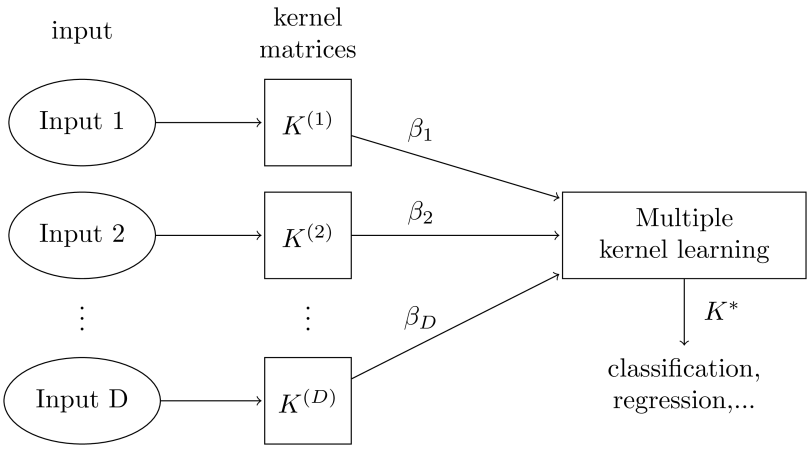



Figure 1: General multiple kernel learning pipeline.

## 2 Multiple kernel learning and logistic regression

In this section we state a binary classification problem by means of a convex combination of multiple kernels. The binary case is considered for simplicity, however the multiclass framework can be treated analogously. Instead of considering a single feature kernel matrix $K$ given by a pre-defined kernel function $k(\cdot,\cdot)$ we propose to resume MKL approaches (see Figure 1). Given a set of training data $\{(x_i, y_i)\}_{i=1}^N$ with $N \in \mathbb{N}$, $x_i$ represents a $P$-dimensional input vector and $y_i \in \{0, 1\}$ its label.

Let $D \in \mathbb{N}$ be the number of different sources: we define a set of kernel matrices $\mathcal{K} := \{K^{(1)}, K^{(2)}, \ldots, K^{(D)}\}$, of sizes $N \times N$, each defined by $K^{(d)} := [k^{(d)}(x_i, x_j)]_{i,j=1}^N$, where $k^{(d)}(\cdot,\cdot)$ is a positive semi-definite kernel function. Note that $K^{(1)}, \ldots, K^{(D)}$ could either arise from different kernel functions (e.g Gaussian, RBF or polynomial kernels) or the same functions with different parameters.
In order to define a linear combination of kernels still being a kernel, let us define a new matrix $K \in \mathbb{R}^{N\times N}$

$$K := \sum_{d=1}^{D} \beta_d K^{(d)} \tag{1}$$

where $\beta_1, \ldots, \beta_D$ are assumed to be non-negative weights. Although in the literature $\beta_1, \ldots, \beta_D$ are usually assumed to take values in the $D-1$ simplex, this property is not needed in order for $K$ to be a kernel and this is why we remove that constraint in a first time. That case, however, will be treated in the long version of this paper.

We define then $K_i^{(d)}$ to be the $i$-th row of the matrix $K^{(d)}$. The $i$-th observation $y_i$ is then described by row vectors $K_i^{(d)}$ for $d \in \{1, \ldots, D\}$. We see $y_i$ as the outcome of a Bernoulli random variable $Y_i$, whose probability of success is denoted by $p_i$. The $N$ random variables

$Y_1, ..., Y_N$ are assumed to be independent (not identically distributed) and

$$\log\left(\frac{p_i}{1-p_i}\right) = \left[\sum_{d=1}^{D} \beta_d K_i^{(d)}\right]\alpha, \tag{2}$$

where $\alpha$ is an unknown vector parameter in $\mathbb{R}^N$. We can rewrite the right-hand side of eq. (2) as

$$\left[\sum_{d=1}^{D} \beta_d K_i^{(d)}\right]\alpha = \sum_{n=1}^{N}\sum_{d=1}^{D} \beta_d K_{in}^{(d)}\alpha_n = \beta^T L_i \alpha,$$

where $\beta := [\beta_1, \ldots, \beta_D]^T \in \mathbb{R}^D$ and $L_i := [K_i^{(1)}, \ldots, K_i^{(D)}]^T \in \mathbb{R}^{D\times N}$ (the $d$-th row of $L_i$ is $K_i^{(d)}$).

The likelihood of the observed data is

$$p(y|\mathcal{K}, \alpha, \beta) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$$

and log-likelihood

$$\log p(y|\mathcal{K}, \alpha, \beta) = \sum_{i=1}^{N}\left[y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right] = \sum_{i=1}^{N}\left[y_i\beta^T L_i\alpha - \log\left(1+e^{\beta^T L_i\alpha}\right)\right]$$

where we used

$$p_i = \frac{e^{\beta^T L_i\alpha}}{1+e^{\beta^T L_i\alpha}}.$$

In order to reduce the risk of over fitting and possibly achieve a better generalization, it is common to consider an $l2$ penalised log-likelihood for the logistic regression model (Cessie and Houwelingen, 1992). One major advantage of such choice is that the log-likelihood remains differentiable and concave and this is why we do adopt that choice too. Now, knowing that, under our assumptions for a given $\beta$, $K$ in eq. (1) is a positive semi-definite kernel matrix, the above log-likelihood is a kernelized one and via the Representer Theorem (Schölkopf et al., 2001) the $l2$ penalty in the feature space translates into $\alpha^T K\alpha$. Then, the penalised log-likelihood is

$$l_R(\alpha, \beta) = \log p(y|\mathcal{K}, \alpha, \beta) - \lambda\alpha^T K\alpha = \sum_{i=1}^{N}\left[y_i\beta^T L_i\alpha - \log\left(1+e^{\beta^T L_i\alpha}\right)\right] - \lambda\alpha^T K\alpha. \tag{3}$$

Optimizing the above quantity jointly in $\alpha$ and $\beta$ is not trivial and no close formula exists, not even when one fixes $\alpha$ (respectively $\beta$) and optimizes with respect to $\beta$ ($\alpha$) although in that case our problem reduces to a standard (kernelized) logistic regression. Instead to directly attack the above problem by numerical maximization, we introduce an additional assumption whose benefits will be clear in a while.

Let us assume that $\beta_1, \ldots, \beta_D \overset{\text{i.i.d.}}{\sim} \text{Exp}(\delta)$ with $\delta$ a fixed positive scalar:

$$p(\beta|\delta) = \prod_{d=1}^{D} p(\beta_d|\delta) = \delta^D e^{-\delta \sum_{d=1}^{D} \beta_d},$$

thus

$$\log p(\beta|\delta) = D \log \delta - \delta \sum_{d=1}^{D} \beta_d = D \log \delta - \delta ||\beta||_1 \tag{4}$$

where the last equality comes from the fact that each $\beta_d$ is non-negative. Thus the log-likelihood of the complete data $(y, \beta)$ reads

$$\log p(y, \beta|\mathcal{K}, \alpha, \delta) = l_R(\alpha, \beta) + \log p(\beta|\delta) = l_R(\alpha, \beta) - \delta \|\beta\|_1 + D \log \delta. \tag{5}$$

As it can be seen a Lasso penalty naturally arises on $\beta$, making this objective function very interesting to optimize since it allow to select the relevant kernels via shrinking. A similar optimization problem was attacked from a pure optimization perspective in (Rakotomamonjy et al., 2008) for SVMs and (He et al., 2021) for logistic regression. However, the interest of the proposed probabilistic formulation is that it opens the doors to stochastic variational inference which allows us to optimize a lower bound of the the log-likelihood of the observed data

$$\log p(y|\mathcal{K}, \alpha, \delta) = \log \int_{\beta} p(y, \beta|\mathcal{K}, \alpha, \delta) d\beta$$

in an incredibly simple way (next section). A final remark before moving to the inference of the model parameters. Similarly to what we did for $\beta$, a prior distribution could be attached to $\alpha$ in order to avoid any estimation and adopt a fully Bayesian perspective, similar to what is done in (Gönen, 2012). However we prefer to maintain the hybrid approach detailed so far in order to keep the model as simple as possible.

# 3 Stochastic variational inference for MKL

Focusing now on the integrated log-likelihood with respect to $\beta$, we have that for any distribution $q(\cdot)$ on $\beta$, with the same support of the prior $p_\delta$, it holds that

$$\begin{aligned}
\log p(y|\mathcal{K}, \alpha) &= \log \int_{\beta} p(y, \beta|\mathcal{K}, \alpha) d\beta \\
&= \log \int_{\beta} \frac{p(y, \beta|\mathcal{K}, \alpha)}{q(\beta)} q(\beta) d\beta \\
&= \log \mathbb{E}_q \left[ \frac{p(y, \beta|\mathcal{K}, \alpha)}{q(\beta)} \right] \geq \mathbb{E}_q \left[ \log \frac{p(y, \beta|\mathcal{K}, \alpha)}{q(\beta)} \right],
\end{aligned} \tag{6}$$

by Jensen inequality. It is also well known that the above inequality turns into an equality when $q(\beta) = p(\beta|y, K, \alpha)$ which is however not analytically tractable here. Thus, we call the lower-bound of eq. (6) $\mathcal{L}(q, \alpha)$ and we set the *variational* (approximate) posterior distribution such as

$$q(\beta) = \prod_{d=1}^{D} q(\beta_d)$$

with $q(\beta_d) := q(\beta_d|\lambda_d)$ assumed to follow an Exponential distribution of parameter $\lambda_d > 0$[1].

By denoting $\lambda := (\lambda_1, \ldots, \lambda_D)$, the lower-bound can be developed as

$$\begin{aligned}
\mathcal{L}(q, \alpha) &= \mathbb{E}_q\big[\log p(y, \beta|\mathcal{K}, \alpha)\big] - \mathbb{E}_q\big[\log q(\beta|\lambda)\big] = \\
&= \mathbb{E}_q\big[l_R(\alpha, \beta)\big] + \mathbb{E}_q\big[\log p(\beta|\delta)\big] - \mathbb{E}_q\big[\log q(\beta|\lambda)\big] = \\
&= \mathbb{E}_q\big[l_R(\alpha, \beta)\big] + \mathbb{E}_q\left[\log \frac{p(\beta|\delta)}{q(\beta|\lambda)}\right] = \\
&= \mathbb{E}_q\big[l_R(\alpha, \beta)\big] - \sum_{d=1}^{D}\left(\frac{\delta}{\lambda_d} + \log \lambda_d\right) + D(\log \delta + 1)
\end{aligned} \tag{7}$$

where the last two terms come from the explicit calculation of the negative Kullback-Leibler divergence between the approximate posterior $q(\cdot|\lambda)$ and the prior distribution $p(\cdot|\delta)$. Since for the Exponential distribution we can adopt the reparametrization trick, i.e.

$$\beta_d = \frac{1}{\lambda_d}\eta,$$

with $\eta \sim$ Exponential(1), following (Kingma and Welling, 2014) (in the context of variational auto-encoders) we can reparametrize the above lower bound in such a way that it is differentiable with respect to the model parameters $\alpha$ and the variational parameters $\lambda$ and obtain an unbiased estimate of its gradient through sampling from independent Exponential distritubions of parameter 1. The model parameters are then optimized by stochastic gradient ascent. More details are in Pseudo-code 1.

**Algorithm 1** SVIMKL

**Require:** $\alpha_0 \in \mathbb{R}^N$, $\ell_0 = [\log D, \ldots, \log D] \in \mathbb{R}^D$
**Ensure:** solution: $(\alpha^*, \lambda^*)$
  $\alpha_c \leftarrow \alpha_0$
  $\ell_c \leftarrow \ell_0$
  **while** $\mathcal{L}(q, \alpha)$ not converged **do**
    $\lambda_c \leftarrow \exp(\ell_c)$
    $\beta_c \leftarrow \frac{1}{\lambda_c}\text{Exp}(1)$
    evaluate $\tilde{\mathcal{L}}(q, \alpha)$ in $\beta_c$ ▷ $\tilde{\mathcal{L}}$: no more $\mathbb{E}_q$ in Eq. (7) and $\beta_c$ in place of $\beta$
    $\alpha_c \leftarrow \alpha_{c+1} = \alpha_c + \nabla_\alpha \tilde{\mathcal{L}}(q, \alpha_c)$
    $\ell_c \leftarrow \ell_{c+1} = \ell_c + \nabla_\ell \tilde{\mathcal{L}}(q, \alpha_c)$
  **end while**
  $(\alpha^*, \lambda^*) \leftarrow (\alpha_c, \lambda_c)$

[1]Variational distributions other than the Exponential are also considered and these choices will discussed at the oral.

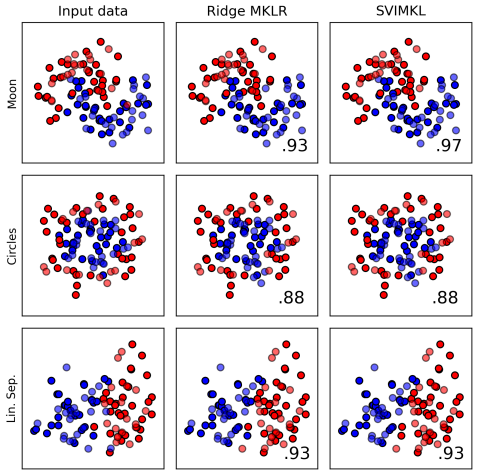



Figure 2: Binary classification results. The bottom right shows the accuracy on the test set.

# 4 Numerical examples

In this section, we present a toy example of binary classification. We consider three 2 dimensional datasets taken from the `sklearn` library: moon, circles and linearly separable data. For each dataset we consider 100 number of points.
In order to perform kernel(s) selection, we consider four ($D = 4$) different kernels. A linear kernel, a sigmoid kernel, an RBF kernel with $\sigma = 0.1$ and a Laplacian kernel with $\sigma = 0.1$. We also compare our method with a multiple kernel logistic regression with a pure optimization approach (MKLR) and an $l2$ penalty on kernel weights.

| Input Data | Ridge MKLR | SVIMKL |
|---|---|---|
| Moon | 1.0244/0.2397/1.1187/3.7014 | 0.0338/0.0899/0.2191/0.6546 |
| Circles | 0.0277/0.0605/1.4507/5.7608 | 0.0964/0.0272/0.1329/0.8220 |
| Lin. sep. data | 1.1034/0.1429/0.8841/2.0121 | 0.2948/0.0853/0.2520/0.1871 |

Table 1: Table of optimal kernel weights ($\beta_1^*$, $\beta_2^*$, $\beta_3^*$, $\beta_4^*$) associated to the linear, sigmoid, RBF and Laplacian kernel respectively.

Table 1 reports the optimal weights associated to the four kernels. In general, our method tends to perform more kernel selection than Ridge MKLR. For example, in the case of the Moon dataset, it can be seen that the weights associated to linear and sigmoid kernels are closer to zero with SVIMKL than those computed with Ridge MKLR. Figure 2 illustrates the accuracy results on the test set.

(More examples will be presented at the oral)

# References


Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Cessie, S. l. and Houwelingen, J. V. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 41(1):191–201.

Damoulas, T. and Girolami, M. A. (2008). Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, 24(10):1264–1270.

Girolami, M. and Rogers, S. (2005). Hierarchicsbayesian models for kernel learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 241–248.

Gönen, M. (2012). Bayesian efficient multiple kernel learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 91–98.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268.

He, X., Huang, J., and Zeng, Z. (2021). Logistic regression based multi-task, multi-kernel learning for emotion recognition. In *2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 572–577. IEEE.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *stat*, 1050:10.

Naesseth, C., Ruiz, F., Linderman, S., and Blei, D. (2017). Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR.

Rakotomamonjy, A., Bach, F., Canu, S., and Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Wang, T., Zhang, L., and Hu, W. (2021). Bridging deep and multiple kernel learning: A review. *Information Fusion*, 67:3–13.

# Classification non supervisée et modèle de mélange

# Kernel KMeans clustering splits for end-to-end unsupervised decision trees

Louis Ohl [1,2] & Pierre-Alexandre Mattei [1] & Mickaël Leclercq [2] & Arnaud Droit [2] & Frédéric Precioso [1]

[1] *Université Côte d'Azur, Inria équipe-projet Maasai, I3S / LJAD, CNRS*
[2] *Université Laval, Centre de recherche du CHU de Québec*
louis.ohl@inria.fr, pierre-alexandre.mattei@inria.fr,
mickael.leclercq@crchudequebec.ulaval.ca,
arnaud.droit@crchudequebec.ulaval.ca, frederic.precioso@univ-cotedazur.fr

**Résumé.** Les arbres de décisions sont des modèles utiles pour obtenir des prédictions avec explications pour des jeux de données de tailles raisonnables. Alors qu'il existe de nombreuses propositions d'algorithmes pour construire de tels arbres en supervisé d'un seul coup, aucune proposition d'algorithme d'apprentissage en une étape n'existe pour le cas non-supervisé du clustering. Les travaux connexes se concentrent plutôt sur l'apprentissage supervisé via un arbre du résultat d'un premier algorithme de clustering. Nous présentons ici une première proposition de construction d'arbre de décisions pour clustering en une seule étape : Kauri. Kauri utilise une optimisation gloutonne du score K-Moyennes à noyau sans calculer ni définir des centroïdes. Nous comparons Kauri à la concaténation K-Moyennes + CART et montrons de meilleurs performances de clustering lorsque le noyau n'est pas linéaire.

**Mot-clefs.** Arbres de décisions, clustering, apprentissage non-supervisé, méthodes de noyaux, K-Moyennes

**Abstract.** Trees are convenient models for obtaining explainable predictions on relatively small datasets. Although there are many proposals for the end-to-end construction of such trees in supervised learning, learning a tree end-to-end for clustering without labels remains an open challenge. As most works focus on interpreting with trees the result of another clustering algorithm, we present here a novel end-to-end trained unsupervised binary tree for clustering: Kauri. This method performs a greedy maximisation of the kernel KMeans objective without requiring the definition of centroids. We compare this model with a KMeans + CART combination and show that Kauri displays better performances for other kernels than the linear kernel.

**Keywords.** Decision trees, clustering, unsupervised learning, kernel methods, KMeans

## 1 Introduction

Decision tree classifiers are one of the most intuitive models in machine learning owing to their intrinsic interpretability (Molnar, 2020, Section 3.2). Trees consist of a set of hierarchically sorted nodes that start from one single root node. Each node comprises two or more

conditions, called rules, each of which leads to a different child node. Once a node does not have any children, a decision is returned. A childless node is named a leaf.

While the end model is eventually interpretable, building it implies some questions to be addressed, notably regarding the number of nodes, the feature (or set of features) on which to apply a decision rule, the construction of a decision rule i.e. the number of thresholds and hence the number of children per node. Learning the structure is easier in the case of supervised learning, whereas the absence of labels makes the construction of unsupervised trees more challenging. In recent related works, the problem was oftentimes addressed with twofold methods (Tavallali et al., 2021; Laber et al., 2023): first learning clusters using another algorithm e.g. KMeans, then applying a supervised decision tree to uncover explanations of the clusters. However, such *unsupervised trees* are not fully unsupervised in fact, since their training still requires the presence of external labels for guidance which are provided by KMeans.

To alleviate this dependence on an exterior clustering algorithm:

- We show how the kernel KMeans objective can be rephrased to avoid the computations of centroids, leveraging simple gains to compute for split proposals in decision trees.
- We then introduce an end-to-end unsupervised tree for clustering: KMeans as unsupervised reward ideal (Kauri), which is not restricted to a fixed number of leaves. To the best of our knowledge, this is the first kernel-based end-to-end clustering tree.
- We show that Kauri often displays better performance in clustering to kernel KMeans + Tree using end-to-end training for kernels other than the linear kernel.
- We finally show that Kauri addresses some limitations of the kernel KMeans algorithm.

# 2 Training trees

In clustering, we do not have access to labels making all notions of gains such as the Gini criterion (Gini, 1912; Breiman, 1984) or the information gain (Quinlan, 1986, 2014) unusable, so we need other tools to guide the decision tree splitting procedure. A common approach is then to keep the algorithm supervised as described in the previous section, yet providing labels that were derived from a clustering algorithm e.g. KMeans (Laber et al., 2023; Held and Buhmann, 1997). In this sense, centroids derived from KMeans can also be involved in split procedures (Tavallali et al., 2021), even to the point that the data from which the centroids are derived do not need to be collected (Gamlath et al., 2021). However, such methods do not properly construct the tree *from scratch* in an unsupervised way despite potential changes in the gain formulations. We are interested in a method that can provide a directly integrated objective to optimise tree training. Other gains derived from entropy formulations have also been proposed (Bock, 1994; Basak and Krishnapuram, 2005). We even note the use of mutual information to achieve deeper and deeper refinements of binary clusters (Karakos et al., 2005).

Oftentimes, these approaches assume that a leaf describes fully a cluster, e.g. Blockeel et al. (1998). Combining leaves into a single cluster requires then post hoc methods (Fraiman

et al., 2013). In such a case, an elegant approach for constructing an unsupervised tree was proposed by Liu et al. (2000) by adding uniform noise to the data and assigning a decision tree to separate the noise from the true data. Such trees put in different leaves dense areas of the data, which can then be labelled manually.

To ensure that several leaves can be assigned to a single cluster, related work also focused on the complete initialisation of a tree and refinement according to a global objective function. For example, Bertsimas et al. (2021) directly maximise the silhouette score or the Dunn index, which are internal clustering metrics and require the initialisation of the tree through greedy construction or KMeans labels. The objective is optimised using a mixed integer optimisation formulation of the tree structure. Lately, Gabidolla and Carreira-Perpinan (2022) proposed to optimise an oblique tree, a structure with logistic regressions at each node, through the alternative optimisation of a distance-based objective, e.g. KMeans, providing pseudo-labels to a tree alternating optimisation problem (Carreira-Perpinan and Tavallali, 2018).

# 3 Kauri: KMeans as unsupervised reward ideal

The Kauri tree is a non-differentiable binary decision tree that looks in many ways alike the CART algorithm. It constructs from scratch a binary tree giving hard clustering assignments to the data by using an objective equivalent to the optimisation of a kernel KMeans. In the Kauri structure, a cluster can be described by several leaves.

## 3.1 Objective function

We consider that we have a dataset of $n$ samples: $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{n}$. We write the partition into $K_{\max}$ clusters as:

$$\mathcal{C}_k = \{\boldsymbol{x}_i \in \mathcal{X}_k\}, \forall k \leq K_{\max}, \tag{1}$$

with $\{\mathcal{X}_k\}_{k=1}^{K_{\max}}$ a partition of the data space $\mathcal{X} \subseteq \mathbb{R}^d$.

The kernel KMeans algorithm minimises the cluster sum of squares in a Hilbert space $\mathcal{H}$ with projection $\varphi$ and kernel $\kappa$ with respect to $K_{\max}$ centroids $\boldsymbol{\mu}_k \dots \boldsymbol{\mu}_{K_{\max}}$:

$$\mathcal{L}_{\text{KMeans}} = \sum_{k=1}^{K_{\max}} \sum_{\boldsymbol{x} \in \mathcal{C}_k} \|\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_k\|_{\mathcal{H}}^2. \tag{2}$$

Instead of computing the sample-wise distance to the centroid using the kernel trick (Dhillon et al., 2004), Kauri optimises this objective without explicitly computing centroids per cluster. To that end, we use the following simple equality:

$$\sum_{\boldsymbol{x} \in \mathcal{C}_k} \|\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_k\|_{\mathcal{H}}^2 = \frac{1}{2\,|\mathcal{C}_k|} \sum_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{C}_k} \|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})\|_{\mathcal{H}}^2. \tag{3}$$

Once inserted into the kernel KMeans objective, we get an alternative formulation without centroids:

$$\mathcal{L}_{\text{KMeans}} = \sum_{k=1}^{K_{\max}} \frac{1}{2\,|\mathcal{C}_k|} \sum_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{C}_k} \|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})\|_{\mathcal{H}}^2. \tag{4}$$

Using the kernel trick, we can rephrase this objective as:

$$\mathcal{L}_{\text{KMeans}} = \sum_{k=1}^{K_{\max}} \frac{1}{2\,|\mathcal{C}_k|} \sum_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{C}_k} (\kappa(\boldsymbol{x},\boldsymbol{x}) + \kappa(\boldsymbol{y},\boldsymbol{y}) - 2\kappa(\boldsymbol{x},\boldsymbol{y})), \tag{5}$$

where the two first kernel terms can be summarised as the size of clusters weighting the diagonal elements of the kernel. Finally, the third term is the grand sum of the kernel of the cluster, and thus:

$$\mathcal{L}_{\text{KMeans}} = \sum_{\boldsymbol{x}\in\mathcal{D}} \kappa(\boldsymbol{x},\boldsymbol{x}) - \sum_{k=1}^{K_{\max}} \frac{1}{|\mathcal{C}_k|} \sum_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{C}_k} \kappa(\boldsymbol{x},\boldsymbol{y}). \tag{6}$$

For the sake of simplicity, we introduce the function $\sigma$ that sums the kernel values $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)\rangle$ of samples indexed by two sets:

$$\sigma(E\times F) = \sum_{\substack{\boldsymbol{x}_i\in E\\ \boldsymbol{x}_j\in F}} \kappa(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{7}$$

We will refer to the $\sigma$ function as the *kernel stock*. This function is bilinear with respect to the input spaces. We provide in Figure 1 a visual explanation of its different usages.

We finally note that the first term of Eq. (6) is a constant because it does not depend on the clustering. With the only the second term remaining, we can remove the minus sign and hence maximise the objective:

$$\mathcal{L}_{\text{Kauri}} = \sum_{k=1}^{K_{\max}} \frac{\sigma(\mathcal{C}_k^2)}{|\mathcal{C}_k|}, \tag{8}$$

Therefore, maximising our objective function $\mathcal{L}$ is equivalent up to a constant to minimising a KMeans objective for any kernel. However, in contrast to Eq. (2), it is not a function of centroids, but a function of a partition.

## 3.2 Tree branching

For supervised trees like CART or ID3, the types of splits are binary and guided by the labels which tell us to which class each child node should go. For unsupervised trees, we must consider all possibilities: to which cluster goes the left child, to which cluster goes the right child on what set of features to do the split, on what threshold in this feature to split and on which nodes. We note $\mathcal{T}_p$ the set of samples reaching the $p$-th node. For a split, let $\mathcal{S}_L$ be the subset of samples from the node samples $\mathcal{T}_p$ that will go to the left child node and $\mathcal{S}_R$ the complementary subset of samples that will go to the right child node. Each child node will be assigned to a different cluster, whether new, already existing, or equal to the parent node's cluster assignment. Let $k_p$ be the current cluster membership of the parent node $p$, $k_L$ the future cluster membership for the left child node and $k_R$ the future cluster membership of the right child node, then $\mathcal{S}_L \cup \mathcal{S}_R = \mathcal{T}_p \subseteq \mathcal{C}_{k_p}$ and after splitting: $\mathcal{S}_L \subseteq \mathcal{C}_{k_L}$ and $\mathcal{S}_R \subseteq \mathcal{C}_{k_R}$.

We enforce the following constraints: a child node must stay in the parent node's cluster if both children leaving would empty the parent's cluster; the creation of a new cluster can only be done under the condition that the number of clusters does not exceed a specified limit $K_{\max}$. We also impose a maximum number of leaves $T_{\max}$ which can be equal to at most the number of samples $n$. It is nonetheless possible that the algorithm stops the splitting procedure if all gains become negative before reaching the maximum number of leaves allowed. Unlike Gabidolla and Carreira-Perpinan (2022) who recently proposed unsupervised oblique trees, we choose to keep splits on a single feature because this lowers the complexity of greedy exploration.

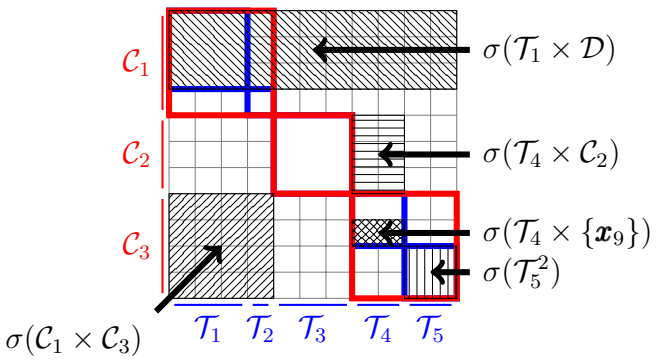



Figure 1: A toy example with a dataset consisting of 11 samples partitioned in 3 clusters using 5 leaves in a tree. The matrix represents the kernel between all pairs of samples and dashed areas correspond to the sum of kernel elements according to the *kernel stock* function $\sigma$.

Thus, learning consists in greedily exploring from all nodes the best split and either taking this split to build a new cluster or merging with another cluster. We now present the objective function and related gains depending on the children's cluster memberships.

## 3.3 Gain metrics

We can derive from the objective of Eq. (8) four gains that evaluate how much score we get by assigning one child node to a new cluster, assigning both child nodes to two new clusters, merging one child node to another cluster, or merging both child nodes to different clusters. We denote by $\mathcal{C}'_\bullet$ the clusters after the split operation and by $\mathcal{C}_\bullet$ the clusters before the split. Hence, the global gain metric is:

$$\Delta\mathcal{L}(\mathcal{S}_L : k_p \to k_L, \mathcal{S}_R : k_p \to k_R) = \frac{\sigma({\mathcal{C}'_{k_L}}^2)}{|\mathcal{C}'_{k_L}|} + \frac{\sigma({\mathcal{C}'_{k_R}}^2)}{|\mathcal{C}'_{k_R}|} + \frac{\sigma({\mathcal{C}'_{k_p}}^2)}{|\mathcal{C}'_{k_p}|} - \frac{\sigma(\mathcal{C}^2_{k_L})}{|\mathcal{C}_{k_L}|} - \frac{\sigma(\mathcal{C}^2_{k_R})}{|\mathcal{C}_{k_R}|} - \frac{\sigma(\mathcal{C}^2_{k_p})}{|\mathcal{C}_{k_p}|}, \quad (9)$$

which corresponds to subtracting the contribution of the kernel stocks of the former clusters and adding the kernel stocks of the new clusters after splitting.

From this global gain metric, we derive four different gains: the *star gain* $\Delta\mathcal{L}^\star$ for assigning either the left or right child of a leaf to a new cluster, the *double star gain* $\Delta\mathcal{L}^{\star\star}$ for assigning the left and right children of a leaf to two new clusters, the *switch gain* $\Delta\mathcal{L}^{\rightleftarrows}$ for assigning either the left or right child of a leaf to another existing cluster and the *reallocation gain* $\Delta\mathcal{L}^{\hookrightarrow}$ for assigning respectively the left and right children to different existing clusters. In

Table 1: Summary of the datasets used in the experiments.

| Name | Breast Cancer | Iris | Lsun | Mice protein | Target | Twodiamonds |
|---|---|---|---|---|---|---|
| Samples | 683 | 150 | 404 | 552 | 770 | 800 |
| Features | 9 | 4 | 3 | 77 | 2 | 2 |
| Classes | 2 | 3 | 3 | 8 | 6 | 2 |

Table 2: ARI scores $_{\text{std}}$ (greater is better) after 30 runs on random subsamples of 80% of the input datasets for varying kernels. All models are limited to finding 4 times more leaves than clusters.

| Kernel | Additive $\chi^2$ | | $\chi^2$ | | Laplacian | | RBF | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Kauri | KMeans+DT | Kauri | KMeans+DT | Kauri | KMeans+DT | Kauri | KMeans+DT |
| Cancer | $0.86_{0.01}$ | $0.84_{0.02}$ | $0.82_{0.02}$ | $0.86_{0.02}$ | $\mathbf{0.87_{0.01}}$ | $0.79_{0.03}$ | $\mathbf{0.87_{0.02}}$ | $0.78_{0.03}$ |
| Iris | $0.67_{0.04}$ | $0.62_{0.09}$ | $0.67_{0.04}$ | $0.61_{0.11}$ | $\mathbf{0.78_{0.05}}$ | $0.71_{0.14}$ | $0.72_{0.03}$ | $0.71_{0.05}$ |
| Lsun | $\mathbf{0.98_{0.00}}$ | $0.89_{0.21}$ | $\mathbf{0.98_{0.01}}$ | $0.81_{0.26}$ | $\mathbf{0.98_{0.01}}$ | $0.96_{0.01}$ | $0.88_{0.02}$ | $0.93_{0.02}$ |
| Twodiamonds | $0.98_{0.00}$ | $0.95_{0.02}$ | $0.98_{0.00}$ | $0.97_{0.02}$ | $\mathbf{1.00_{0.00}}$ | $0.77_{0.43}$ | $\mathbf{1.00_{0.00}}$ | $\mathbf{1.00_{0.00}}$ |
| Wine | $0.87_{0.03}$ | $0.74_{0.03}$ | $\mathbf{0.90_{0.03}}$ | $0.73_{0.10}$ | $0.89_{0.03}$ | $0.83_{0.03}$ | $0.85_{0.03}$ | $0.80_{0.04}$ |

practice, for the $p$-th leaf with split proposals $\mathcal{S}_L$ and $\mathcal{S}_R$, we do not need to compute the future cluster kernel stocks $\sigma(\mathcal{C}'_\bullet)$ and instead use the stocks between the splits and the current clusters: $\sigma(\mathcal{S}_{L/R} \times \mathcal{C}_\bullet)$. The code of Kauri is available in the GemClus package: `https://gemini-clustering.github.io`.

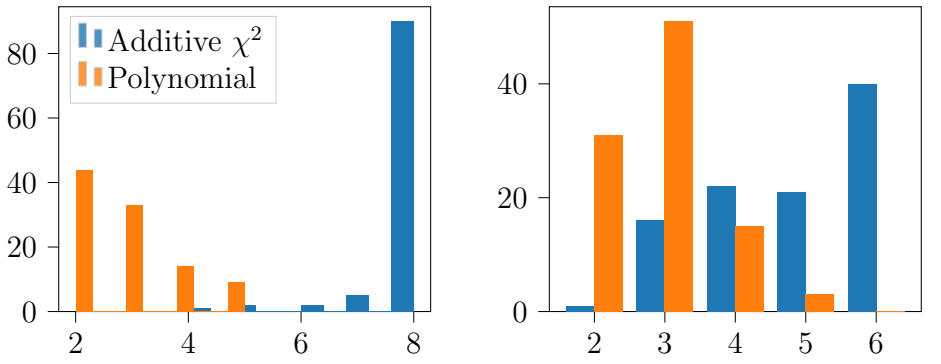
# 4 Experiments

The datasets used in the experiments are described in Table 1. We will assess the general clustering performances using the adjusted rand index (ARI, Hubert and Arabie, 1985) and the KMeans score normalised by the reference score of the sole kernel KMeans algorithm for cluster quality. We used minmax scaling for all datasets to ensure the computation of $\chi^2$ kernels. We tossed away all samples that presented missing values for the sake of simplicity. Finally, we introduced stochasticity by measuring the performances on random samples of 80% of the dataset to be fair between the stochastic nature of kernel KMeans and the deterministic nature of Kauri due to its greedy construction.

We compare the performances of Kauri against the concatenation of a kernel KMeans algorithm and a supervised decision tree for different kernel (KMeans+DT). However, using a different kernel leads to the absence of a definition of centroids in the Euclidean space where the data lie. Consequently, the available implementations of the related work (Frost et al., 2020; Makarychev and Shan, 2022; Laber et al., 2023) are not compatible with this set-up because they generally require a centroid. We explore 4 different kernels with default parameters from `scikit-learn` (Pedregosa et al., 2011): $\chi^2$, additive $\chi^2$, Laplacian and RBF kernels with Table 2 for the ARI and Table 3 for the normalised KMeans score. We also allow the algorithms to find 4 times more leaves than clusters, as was done by Frost et al. (2020).

Table 3: Relative Kernel KMeans scores $_{\text{std}}$ (lower is better) of Kauri and Kernel-KMeans + Decision Tree after 30 runs on random subsamples of 80% of the input datasets for varying kernels. All models are limited to finding 4 times more leaves than clusters.

| Kernel | Additive $\chi^2$ | | $\chi^2$ | | Laplacian | | RBF | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Kauri | KMeans+DT | Kauri | KMeans+DT | Kauri | KMeans+DT | Kauri | KMeans+DT |
| Cancer | $1.04_{0.02}$ | $1.05_{0.02}$ | $1.02_{0.01}$ | $1.03_{0.01}$ | $\mathbf{1.01_{0.02}}$ | $1.04_{0.02}$ | $1.04_{0.02}$ | $1.07_{0.03}$ |
| Iris | $1.11_{0.05}$ | $1.18_{0.16}$ | $1.10_{0.04}$ | $1.20_{0.16}$ | $\mathbf{1.05_{0.02}}$ | $1.09_{0.08}$ | $1.06_{0.05}$ | $1.06_{0.09}$ |
| Lsun | $1.08_{0.03}$ | $1.13_{0.11}$ | $\mathbf{1.04_{0.02}}$ | $1.11_{0.12}$ | $\mathbf{1.04_{0.01}}$ | $\mathbf{1.04_{0.01}}$ | $1.05_{0.03}$ | $1.07_{0.04}$ |
| Twodiamonds | $1.03_{0.02}$ | $1.03_{0.02}$ | $1.05_{0.02}$ | $1.05_{0.02}$ | $\mathbf{1.01_{0.01}}$ | $1.06_{0.09}$ | $1.04_{0.01}$ | $1.03_{0.02}$ |
| Wine | $1.03_{0.03}$ | $1.05_{0.04}$ | $1.05_{0.02}$ | $1.08_{0.05}$ | $\mathbf{1.02_{0.02}}$ | $\mathbf{1.02_{0.02}}$ | $1.06_{0.03}$ | $1.08_{0.03}$ |




(a) Mice protein (8 classes)

(b) Target (6 classes)

Figure 2: Number of non-empty clusters for 100 runs of kernel KMeans with an additive $\chi^2$ or polynomial kernel. The algorithm had to find the same number of clusters as classes per dataset.

We generally observe equal or stronger ARI scores for the KAURI algorithm in Table 2. The same observation goes for the KMeans score in Table 3, especially for the Laplacian kernel in both tables. Note that we could not run this experiment on the Mice protein and Target datasets. This is due to the phenomenon of empty clusters that arises in the kernel KMeans algorithm. Consequently, the basis on which the decision tree is learnt does not provide enough clusters, thus lowering the ARI and increasing the kernel KMeans score for these two excluded datasets. Similarly, the reference kernel KMeans score used for normalisation suffered from the same problem. Empirically, we observed this behaviour for 2 implementations of the kernel KMeans algorithm.[1]

As a simple example, we ran 100 kernel KMeans models with an additive $\chi^2$ kernel or a polynomial kernel for the Mice protein dataset and the Target dataset. From Figure 2, we observe that for a relatively small number of clusters, the kernel KMeans may converge to several empty clusters, with the worst effect from the polynomial kernel in this example. To complete Figure 2, the number of clusters found with the best kernel KMeans score was 8

[1] `https://gist.github.com/mblondel/6230787`, and the `tslearn` v0.6.3 implementation (Tavenard et al., 2020).

for the Mice protein dataset and 6 for the Target dataset. In contrast, datasets with only 2 clusters to find often converged to the good number of clusters, making scores comparable.

This shows that with the end-to-end construction of the Kauri tree, we do not suffer from dependence to the basis KMeans algorithm, and manage to get the correct user-desired number of clusters.

# 5 Final words

We introduced a novel model optimising a kernel KMeans objective: Kauri. By rephrasing the kernel KMeans objective to drop the requirement for centroids, we leveraged an easy criterion to maximise and derived gains for an iterative splitting procedure in a CART-like binary tree. In contrast to related work on unsupervised binary decision trees, Kauri is compatible with kernels other than the linear kernel. Moreover, we showed that the algorithm does not suffer from an empty cluster phenomenon that arises in the kernel KMeans algorithm and will fill all clusters. When the kernel KMeans algorithm converges to the correct number of clusters, the performances of Kauri remain greater than the KMeans+DT baseline. Hence, the strong advantage of this method is building an interpretable by-nature clustering instead of seeking to explain another clustering output from a different algorithm. Future work will focus on the integration of linear splits using several features, such as in oblique trees (Gabidolla and Carreira-Perpinan, 2022).

## Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur, Investment in the Future, project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. We would also like to thank the France Canada Research Fund (FFCR) for their contribution to the project. This work was partly supported by the Fonds de recherche du Québec – Santé (FRQS) and the Health-Data Hub, through the joint project AORTIC STENOSIS.

# References

Jayanta Basak and Raghu Krishnapuram. Interpretable Hierarchical Clustering by Constructing an Unsupervised Decision Tree. *IEEE transactions on knowledge and data engineering*, 17(1):121–132, 2005. Publisher: IEEE.

Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable Clustering: an Optimization Approach. *Machine Learning*, 110(1):89–138, January 2021. ISSN 1573-0565. doi: 10.1007/s10994-020-05896-2.

Hendrik Blockeel, Luc De Raedt, and Jan Ramon. Top-Down Induction of Clustering Trees. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 55–63, 1998.

HH Bock. Information and Entropy in Cluster Analysis. In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach: Volume 2 Multivariate Statistical Modeling*, pages 115–147. Springer, 1994.

L Breiman. Classification and Regression Trees. *The Wadsworth & Brooks/Cole*, 1984. Publisher: Advanced Books & Software.

Miguel A Carreira-Perpinan and Pooya Tavallali. Alternating optimization of decision trees, with application to learning sparse oblique trees. *Advances in neural information processing systems*, 31, 2018.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel K-Means: Spectral Clustering and Normalized Cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, 2004.

Ricardo Fraiman, Badih Ghattas, and Marcela Svarc. Interpretable Clustering using Unsupervised Binary Trees. *Advances in Data Analysis and Classification*, 7:125–145, 2013. Publisher: Springer.

Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. ExKMC: Expanding Explainable k-Means Clustering. *arXiv preprint arXiv:2006.02399*, 2020.

Magzhan Gabidolla and Miguel A Carreira-Perpinan. Optimal Interpretable Clustering using Oblique Decision Trees. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 400–410, 2022.

Buddhima Gamlath, Xinrui Jia, Adam Polak, and Ola Svensson. Nearly-Tight and Oblivious Algorithms for Explainable Clustering. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28929–28939. Curran Associates, Inc., 2021.

Corrado W Gini. Variability and Mutability, Contribution to the Study of Statistical Distributions and Relations. *Studi Economico-Giuridici della R. Universita de Cagliari*, 1912.

Marcus Held and Joachim Buhmann. Unsupervised on-line Learning of Decision Trees for Hierarchical Data Analysis. *Advances in neural information processing systems*, 10, 1997.

Lawrence Hubert and Phipps Arabie. Comparing Partitions. *Journal of classification*, 2(1): 193–218, 1985. Publisher: Springer.

Damianos Karakos, Sanjeev Khudanpur, Jason Eisner, and Carey E Priebe. Unsupervised Classification via Decision Trees: An Information-Theoretic Perspective. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–1081. IEEE, 2005.

Eduardo Laber, Lucas Murtinho, and Felipe Oliveira. Shallow Decision Trees for Explainable K-means Clustering. *Pattern Recognition*, 137:109239, 2023. Publisher: Elsevier.

Bing Liu, Yiyuan Xia, and Philip S Yu. Clustering Through Decision Tree Construction. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 20–29, 2000.

Konstantin Makarychev and Liren Shan. Explainable K-Means: Don't be Greedy, Plant Bigger Trees! In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, pages 1629–1642, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9264-8. doi: 10.1145/3519935.3520056. event-place: Rome, Italy.

Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Ross Quinlan. Induction of Decision Trees. *Machine learning*, 1:81–106, 1986. Publisher: Springer.

J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

Pooya Tavallali, Peyman Tavallali, and Mukesh Singhal. K-means Tree: An Optimal Clustering Tree for Unsupervised Learning. *The Journal of Supercomputing*, 77:5239–5266, 2021. Publisher: Springer.

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. Tslearn, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research*, 21(118):1–6, 2020.

# ESTIMATION PARAMÉTRIQUE D'UN MODÈLE $q$-GAUSSIEN ET D'UN MODÈLE MÉLANGE DE $q$-GAUSSIENNES

Oumaima Ben Mrad[1] & Afif Masmoudi[2]& Yousri Slaoui [3]

[1] *Laboratoire de Probabilités et Statistique, Sfax, Tunisie et Laboratoire de Mathématiques et Applications, Poitiers, France. oumaima.benmrad.fss@gmail.com*
[2] *Laboratoire de Probabilités et Statistique, Sfax, Tunisie. afif.masmoudi@fss.usf.tn*
[3] *Laboratoire de Mathématiques et Applications, Poitiers, France. Yousri.Slaoui@math.univ-poitiers.fr*

**Résumé.** Dans ce travail, nous abordons l'estimation des paramètres de la $q$-distribution $q$-Gaussienne $\mathscr{N}_q(\mu,\sigma^2)$ introduite par Ben Mrad et al. (2023), une généralisation de la $q$-distribution $\mathscr{N}_q(0,1)$ introduite par Diaz et Pariguan (2009). Pour ce faire, nous commençons par introduire une nouvelle distribution discrète, appelée $q$-Gaussienne discrète $\mathscr{N}_q^d(\mu,\sigma^2)$, associée à chaque $q$-Gaussienne quantique introduite par Diaz et Pariguan (2009). Ensuite, nous procédons à l'estimation des paramètres de la distribution $\mathscr{N}_q^d(\mu,\sigma^2)$, ce qui implique une estimation des paramètres de la distribution $q$-Gaussienne $\mathscr{N}_q(\mu,\sigma^2)$. Nous appliquons pour cela les méthodes des moments et du maximum de vraisemblance. De plus, nous étudions un mélange fini de $q$-Gaussiennes discrètes et appliquons l'algorithme Espérance-Maximisation (E-M) pour estimer les paramètres du mélange. Enfin, nous menons une étude de simulation pour évaluer le modèle et les méthodes d'estimation.

**Mots-clés.** $q$-Calcul, $q$-Distribution, $q$-Gaussienne, Mélange fini, Estimation Paramétrique, Algorithme E-M.

**Abstract.** In this work, we address parameters estimation of the $q$-Gaussian $q$-distribution $\mathscr{N}_q(\mu,\sigma^2)$ introduced by Ben Mrad et al. (2023), a generalization of the $q$-distribution $\mathscr{N}_q(0,1)$ introduced by Diaz and Pariguan (2009). To do this, we begin by introducing a new discrete distribution, called $q$-discrete Gaussian $\mathscr{N}_q^d(\mu,\sigma^2)$, associated for each quantum $q$-Gaussian introduced by Diaz and Pariguan (2009). Next, we proceed to estimate the parameters of $\mathscr{N}_q^d(\mu,\sigma^2)$ distribution, which involves estimating the parameters of the $q$-Gaussian $q$-distribution $mathcalN_q(\mu,\sigma^2)$. We apply the methods of moments and maximum likelihood. Then, we study a finite mixture of $q$-discrete Gaussians and apply the Expectation-Maximization (E-M) algorithm to estimate the mixture parameters. Finally, we conduct a simulation study to evaluate the model and estimation methods.

**Keywords.** $q$-Calculus, $q$-Distribution, $q$-Gaussian, Finite mixture, Parametric estimation, E-M algorithm.

# 1 Introduction

La naissance du calcul quantique, également connu sous le nom de $q$-calcul, remonte à Léonard Euler, qui a introduit le paramètre $q$ dans les séries infinies de Newton en 1740. Depuis sa création et jusqu'à nos jours, le $q$-calcul est un sujet de recherche et de développement majeur dans plusieurs domaines mathématiques et physiques. Ensuite, en 1910, Jackson (1910) a introduit les concepts de la $q$-intégrale et de la $q$-dérivée, ce qui a été une contribution marquante. Dans leur livre, Kac et Cheung (2002) ont repris et amélioré plus tard le $q$-calcul.

En analyse quantique, on s'intéresse généralement aux $q$-analogues, où il n'existe pas une règle générale pour définir la notion du $q-$analogue. Une définition intuitive d'un $q$-analogue d'un tel objet mathématique $A$, est une famille d'objets $(A_q)_{q\in\mathbb{R}\setminus\{1\}}$, tels que $A_q$ tend vers $A$ lorsque $q$ tend vers 1. En particulier, plusieurs mathématiciens ont été intéressés par la construction d'un $q$-analogue de la distribution Gaussienne classique qui revient à la construction d'un q-analogue de sa fonction densité, appelée fonction $q$-densité. Parmis eux on trouve Diaz et Pariguan qui ont introduit un $q$-analogue de la distribution Gaussienne centrée réduite notée $\mathcal{N}_q(0,1)$. La fonction $q$-densité de cette $q$-distribution est donnée par

$$s_q(x) = \frac{1}{c(q)} E_{q^2}^{-\frac{q^2x^2}{[2]_q}} \mathbb{1}_{[-\nu,\nu]}(x) \ \forall 0 < q < 1. \tag{1.1}$$

Où $c(q) := \int_{-\nu}^{\nu} E_{q^2}^{-\frac{q^2x^2}{[2]_q}} d_q x$ est la constante de normalisation de $\mathcal{N}_q(0,1)$, $\nu := \frac{1}{\sqrt{1-q}}$, $E_q^x$ est la fonction $q$-exponentielle introduite par Jackson (2010) définie par $E_q^x = \sum_{j=0}^{\infty} q^{j(j-1)/2} \frac{x^j}{[j]_q!}$. Avec, $[\alpha]_q = \frac{1-q^\alpha}{1-q}$, $\forall \alpha \in \mathbb{R}$ et $[n]_q! = \begin{cases} 1 & si \ \ n = 0 \\ [n]_q[n-1]_q...[1]_q & si \ n \geq 1. \end{cases}$

Dans notre papier, Ben Mrad et al. (2023), nous avons généralisé cette $q$-Gaussienne en une $q$-Gaussienne de moyenne $\mu$ et de variance $\sigma^2$ de la façon suivante : En premier lieu, nous avons introduit la $q$-distribution $\mathcal{N}_q(0,\sigma^2)$, de fonction $q$-densité définie par

$$g_q(y) = \frac{1}{c(\sigma,q)} E_{q^2}^{-\frac{q^2y^2}{\sigma^2[2]_q}} \mathbb{1}_{[-\sigma\nu,\sigma\nu]}(y).$$

En deuxième lieu, nous avons caractérisé la $q$-Gaussienne $\mathcal{N}_q(\mu,\sigma^2)$ par :

$$Z \sim \mathcal{N}_q(\mu,\sigma^2), \ \text{ si } \ Z-\mu \sim \mathcal{N}_q(0,\sigma^2).$$

Notre objectif dans cette communication est d'estimer les paramètres $q$, $\mu$ et $\sigma^2$ de la $q$-Gaussienne $\mathcal{N}_q(\mu,\sigma^2)$ et d'étudier et d'estimer un modèle de mélange fini de $q$-Gaussiennes. Les résultats présentés dans cette communication fait l'objet d'un article intitulé "The discrete $q$-Gaussian distribution $\mathcal{N}_q(\mu,\sigma^2)$ : Properties and parameters estimation" publié dans le journal "Pysics Letters A" (voir Ben Mrad et al. (2024)).

## 1.1 La distribution $q$-Gaussienne discrète $\mathcal{N}_q^d(\mu,\sigma^2)$

La définition suivante présente une nouvelle distribution discrète, appelée $q$-Gaussienne discrète centrée réduite $\mathcal{N}_q^d(0,1)$, associée à chaque $q$-Gaussienne quantique introduite par Diaz et Pariguan (2009) donnée par l'Equation 1.1.

**Définition 1.1.** *Soient* $0<q<1$, $\nu=\dfrac{1}{\sqrt{1-q}}$ *et* $T:=\{\pm q^j\nu; j\geq 0\}\bigcup\{0\}$. *On dit qu'une variable aléatoire X de support T suit la distribution discrète q-Gaussienne, notée* $\mathcal{N}_q^d(0,1)$, *si sa fonction de masse est donnée par*

$$\mathbb{P}(X=x)=(1-q)|x|s_q(x)\mathbb{1}_T(x).$$

Cette distribution peut être généraliser en une distribution discrète $q$-Gaussienne de moyenne $\mu$ et d'écart type $\sigma$ en faisant un changement de variable affine à partir d'une variable aléatoire $X\sim\mathcal{N}_q^d(0,1)$, comme le montre la proposition suivante.

**Proposition 1.1.** *Soient* $\mu\in\mathbb{R}$, $\sigma^2>0$ *et* $X\sim\mathcal{N}_q^d(0,1)$. *Alors, la variable aléatoire* $Y:=\sigma X+\mu$ *suit la distribution discrète q-Gaussienne de moyenne* $\mu$ *et d'écart type* $\sigma$, *notée* $Y\sim\mathcal{N}_q^d(\mu,\sigma^2)$, *et sa fonction de masse est définie comme suit*

$$\mathbb{P}(Y=y)=\frac{1-q}{c(q)}\left|\frac{y-\mu}{\sigma}\right|E_{q^2}^{-\frac{q^2(y-\mu)^2}{[2]_q\sigma^2}}\mathbb{1}_T\left(\frac{y-\mu}{\sigma}\right). \tag{1.2}$$

**Théorème 1.1.** *Soient* $\mu\in\mathbb{R}$, $\sigma^2>0$ *et* $0<q<1$. *Alors, on a*

$$\mathcal{U}_{\{-\sigma+\mu;\sigma+\mu\}}\xleftarrow[q\to 0]{\mathscr{L}}\mathcal{N}_q^d(\mu,\sigma^2)\xrightarrow[q\to 1]{\mathscr{L}}\mathcal{N}(\mu,\sigma^2).$$

*Où* $\mathscr{L}$ *se réfère à la convergence en loi et* $\mathcal{U}_{\{-\sigma+\mu;\sigma+\mu\}}$ *se réfère à la distribution Uniforme discrète classique sur* $\{-\sigma+\mu;\sigma+\mu\}$.

Le théorème suivant illustre l'expression des moments de la distribution proposée.

**Théorème 1.2.** *Soit* $Y\sim\mathcal{N}_q^d(\mu,\sigma^2)$. *Alors, les moments de la variable aléatoire Y sont donnés par*

$$\mathbb{E}(Y^n)=\mu^n+\sum_{\substack{k=2,\\ k\ pair}}^{n}\begin{bmatrix}n\\k\end{bmatrix}\sigma^k\mu^{n-k}[k-1]_q!!.$$

*Où,*

$$[n]_q!!=\begin{cases}\displaystyle\prod_{k=1}^{\frac{n}{2}}[2k]_q=[n]_q[n-2]_q\ldots[4]_q[2]_q & si\ \ n\ \ est\ pair\\ \displaystyle\prod_{k=1}^{\frac{n+1}{2}}[2k-1]_q=[n]_q[n-2]_q\ldots[3]_q[1]_q & si\ \ n\ \ est\ impair\end{cases}$$

*et* $\begin{bmatrix}n\\k\end{bmatrix}=\dfrac{n!}{k!(n-k)!}$ *se réfère au coefficient binomial classique et* $n!$ *se réfère à la factorielle de n.*

En se basant sur le papier de Ben Mrad et al. (2023), nous remarquons que les moments de la distribution $q$-Gaussienne discrète $\mathscr{N}_q^d(\mu,\sigma^2)$ et les moments de la $q$-distribution $q$-Gaussienne $\mathscr{N}_q(\mu,\sigma^2)$ sont égaux. En conséquence, estimer les paramètres de $\mathscr{N}_q(\mu,\sigma^2)$ revient à estimer les paramètres de $\mathscr{N}_q^d(\mu,\sigma^2)$.

# 2 Estimation paramétrique de $\mathscr{N}_q^d(\mu,\sigma^2)$

Soient $(x_1,x_2,...,x_n)$ $n$ observations d'un échantillon de variables aléatoires indépendantes et identiquement distribuées suivant la distribution $\mathscr{N}_q^d(\mu,\sigma^2)$ de même fonction de masse donnée par

$$P(X=x)=\frac{1-q}{c(q)}\left|\frac{x-\mu}{\sigma}\right|E_{q^2}^{-\frac{q^2(x-\mu)^2}{[2]_q\sigma^2}}\mathbb{1}_T\left(\frac{x-\mu}{\sigma}\right).$$

L'estimation par la méthode des moments est donnée dans la sous section suivante.

## 2.1 Estimation paramétrique de $\mathscr{N}_q^d(\mu,\sigma^2)$ par la méthode des moments

Dans ce cas, nous supposons que $\mu$, $\sigma^2$ et $q$ sont des paramètres inconnus et nous allons les estimer en utilisant la méthode des moments (MM) comme suit

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i:=\bar{x}.$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

En approximant le moment théorique centré réduit d'ordre 4 par le moment empirique centré réduit d'ordre 4, nous obtenons

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\hat{\sigma}}\right)^4=\mathbb{E}\left(\left(\frac{X-\mu}{\sigma}\right)^4\right).$$

Alors, nous trouvons

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\hat{\sigma}}\right)^4=[3]_q!!=\frac{1-q^3}{1-q}.$$

En résolvant cette équation, nous obtenons un estimateur de $q\in]0,1[$ par la méthode des moments qui est donné par

$$\hat{q}=\frac{-1+\sqrt{1-4\left(1-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\hat{\sigma}}\right)^4\right)}}{2}. \tag{2.1}$$

La sous section suivante décrit en détail l'estimation des paramètres par la méthode du maximum de vraisemblance.

## 2.2 Estimation de $\mu$ et $\sigma^2$ par la méthode du maximum de vraisemblance

Dans ce cas, nous supposons que $q$ est un paramètre connu et on note par $\Theta = (\mu, \sigma^2)$ l'ensemble des paramètres inconnus. Notre but est d'estimer $\Theta$ par la méthode du maximum de vraisemblance.

Les fonctions strictement concave vraisemblance et log-vraisemblance associées au vecteur des paramètres $\Theta = (\mu, \sigma^2)$ sont données, respectivement, par

$$l(x_1, x_2, ...x_n; \Theta) = \prod_{i=1}^{n} \frac{(1-q)}{c(q)} \left| \frac{x_i - \mu}{\sigma} \right| E_{q^2}^{-\frac{q^2(x_i-\mu)^2}{\sigma^2 [2]_q}}.$$

$$L(x_1, x_2, ...x_n; \Theta) = \sum_{i=1}^{n} log(1-q) - log\,(c(q)) + log\,|x_i - \mu| - log(\sigma) + log \left( E_{q^2}^{-\frac{q^2(x_i-\mu)^2}{\sigma^2 [2]_q}} \right).$$

les dérivées partielle de $L(x_1, x_2, ...x_n; \Theta)$ sont données par

$$\begin{cases} \dfrac{\partial L}{\partial \mu} & = \displaystyle\sum_{i=1}^{n} \frac{-1}{x_i - \mu} + \sum_{l \geq 0} \frac{2q^{2l+2}(x_i - \mu)(1-q)t}{1-(1-q)q^{2l+2}(x_i-\mu)^2 t} \\ \dfrac{\partial L}{\partial t} & = \dfrac{n}{2t} + \displaystyle\sum_{i=1}^{n} \sum_{l \geq 0} \frac{-q^{2l+2}(x_i-\mu)^2(1-q)}{1-(1-q)q^{2l+2}(x_i-\mu)^2 t}. \end{cases} \tag{2.2}$$

Par conséquence, l'estimateur du maximum de vraisemblance de $\Theta$ est obtenu en résolvant le système 2.2 comme étant égal à 0.
Dans la sous section suivante, nous examinons un mélange fini de distributions discrètes $q$-Gaussiennes et nous estimons les paramètres du mélange en appliquant l'algorithme (E-M).

## 2.3 Estimation paramétrique d'un mélange fini de $q$-Gaussiennes discrète

Soient $\mu \in \mathbb{R}$, $\sigma^2 > 0$ et $X$ une variable aléatoire suivant la distribution $\mathcal{N}_q^d(\mu, \sigma^2)$. La fonction de masse de $X$ est donnée par

$$\mathbb{P}(x; \mu, \sigma^2, q) = \frac{1-q}{c(q)} \left| \frac{x-\mu}{\sigma} \right| E_{q^2}^{-\frac{q^2 \left(\frac{x-\mu}{\sigma}\right)^2}{[2]_q}} \mathbb{1}_T \left( \frac{x-\mu}{\sigma} \right).$$

La fonction de masse du mélange de $q$-Gaussiennes discrètes est donnée par

$$\mathbb{P}(x \mid \Theta) = \sum_{k=1}^{K} \pi_k \frac{1-q}{c(q)} \left| \frac{x-\mu_k}{\sigma_k} \right| E_{q^2}^{-\frac{q^2 \left(\frac{x-\mu_k}{\sigma_k}\right)^2}{[2]_q}} \mathbb{1}_T \left( \frac{x-\mu_k}{\sigma_k} \right). \tag{2.3}$$

Où $\Theta = \left(\pi_1, ..., \pi_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2, q\right)$ représente le vecteur des paramètres du modèle de mélange, $\theta_k := \left\{\mu_k, \sigma_k^2, q\right\}$ est l'ensemble des paramètres pour le $k^{ème}$ composant du mélange et $\pi_1, \pi_2, ..., \pi_K$ représentent les poids du mélange tels que $(0 < \pi_k < 1$ et $\sum_{k=1}^{K} \pi_k = 1)$. Nous nous intéressons actuellement à l'estimation du vecteur des paramètres du modèle mélange. Nous avons proposé d'appliquer l'algorithme Espérance-Maximisation (E-M) introduit par Dempster et al. (1977) pour estimer le vecteur des paramètres $\left\{\mu_k, \sigma_k^2; 1 \leq k \leq K\right\}$ et la méthode des moments pour estimer le paramètre commun inconnu $q$. Pour cela, nous considérons $(X_1, X_2, ..., X_N)$ comme étant $N$ variables aléatoires indépendantes tirées du mélange donné par l'Équation (2.3), et $(x_1, x_2, ..., x_N)$ comme $N$ observations associées à $(X_1, X_2, ..., X_N)$. La fonction de vraisemblance incomplète est donnée par

$$l(x_1, x_2, ..., x_N; \Theta) = \prod_{i=1}^{N} \mathbb{P}(x_i \mid \Theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \frac{1-q}{c(q)} \left|\frac{x_i - \mu_k}{\sigma_k}\right| E_{q^2}^{-\frac{q^2 \left(\frac{x_i - \mu_k}{\sigma_k}\right)^2}{[2]_q}} \mathbb{1}_T\left(\frac{x_i - \mu_k}{\sigma_k}\right).$$

Afin d'estimer le vecteur des paramètres $\Theta$, nous associons à chaque point de données observé $X_i$ un vecteur aléatoire discret latent $Z_i = (Z_{i1}, ..., Z_{iK})$. Ce vecteur aléatoire latent est modélisé par une distribution de Bernoulli multivariée avec un vecteur de probabilités $(\pi_1, ..., \pi_K)$, c'est-à-dire $\mathbb{P}(Z_i = z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}$ où $z_i = (z_{i1}, ..., z_{iK}) \in \{0,1\}^K$ et $\sum_{k=1}^{K} z_{ik} = 1$. La fonction vraisemblance complètes est donc donnée par

$$l_c(x_1, ..., x_n, z_1, ..., z_n \mid \Theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\pi_k \mathbb{P}_k(x_i \mid \theta_k)\right)^{z_{ik}}.$$

Ainsi, la log-vraisemblance des données complètes s'écrit comme suit :

$$L_c(x_1, ..., x_n, z_1, ..., z_n \mid \Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log(\pi_k) + \log(\mathbb{P}_k(x_i \mid \theta_k))\right].$$

En raison de la présence des données manquantes, nous suggérons, dans notre travail, d'estimer les paramètres par l'algorithme (E-M), qui alterne deux étapes lorsqu'on donne les paramètres initiaux $\Theta^{(0)}$ :

- **Étape Espérance :** L'espérance conditionnelle de la fonction log-vraisemblance des données complètes sachant les observations et le vecteur $\Theta^{(l)}$, qui représente le vecteur des paramètres à la $l^{me}$ itération, est donnée par

$$\begin{aligned} Q(\Theta \mid\mid \Theta^{(l)}) &= \mathbb{E}_{\Theta^{(l)}}(L_c(x_1, ..., x_n, z_1, ..., z_n \mid \Theta) \mid x_1, ..., x_n) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{\Theta^{(l)}}(z_{ik} \mid x_1, ...x_n) \left[\log(\pi_k) + \log(\mathbb{P}_k(x_i \mid \theta_k))\right]. \end{aligned}$$

Nous notons $\tau_{ik}^{(l)} := \mathbb{E}_{\Theta^{(l)}}(z_{ik} \mid x_1, ....x_n) = \dfrac{\pi_k^{(l)} \mathbb{P}_K(x_i; \theta_k^{(l)})}{\sum_{k=1}^{K} \pi_k^{(l)} \mathbb{P}_K(x_i; \theta_k^{(l)})}$, la probabilité à postériori que $X_i$ appartienne à la $k^{ème}$ composante du mélange à la $l^{th}$ itération. L'expression de la quantité intermédiaire $Q(\Theta \mid\mid \Theta^{(l)})$ devienne alors

$$Q(\Theta \mid\mid \Theta^{(l)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(l)} \left[\log(\pi_k \mathbb{P}_k(x_i \mid \theta_k))\right].$$

- **Étape Maximisation :** Il s'agit de maximiser $Q(\Theta \mid\mid \Theta^{(l)})$ globalement par rapport à $\Theta$.

$$\Theta^{(l+1)} = \underset{\Theta}{Argmax}\, Q(\Theta \mid\mid \Theta^{(l)}).$$

Pour $\varepsilon > 0$ choisie par l'utilisateur assez proche de zéro, et une norme sélectionnée sur l'espace des paramètres, l'algorithme (E-M) s'arrête lorsque la norme de la différence entre deux estimations successives des paramètres est suffisamment petite, c'est-à-dire $\parallel \Theta^{(l+1)} - \Theta^{(l)} \parallel \leq \varepsilon$.

**Théorème 2.1.** *Pour $t_k = \dfrac{1}{\sigma_k^2}$, l'algorithme (E-M) à l'itération $(l+1)$ donne les résultats suivants.*

1. *L'estimateur de la proportion $\pi_k$ du mélange est donné par*

$$\pi_k^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(l)}.$$

2. *Les paramètres $\mu_k$ et $\sigma_k^2$ de la $k^{ème}$ composante du mélange peuvent être trouvés en résolvant les équations suivantes :*

$$\sum_{i=0}^{n} \tau_{ik}^{(l)} \frac{\partial \log(\mathbb{P}_k(x_i \mid \mu_k, t_k))}{\partial \mu_k} = 0 \;\; et \;\; \sum_{i=0}^{n} \tau_{ik}^{(l)} \frac{\partial \log(\mathbb{P}_k(x_i \mid \mu_k, t_k))}{\partial t_k} = 0.$$

La section suivante présente des études de simulation, qui ont pour but d'évaluer la performance des estimateurs obtenus.

# 3 Étude numérique

En utilisant la technique classique de simulation par inversion de la fonction de répartition, un échantillon de taille $N = 5000$ est simulé à partir de la distribution discrète $q$-Gaussienne $\mathcal{N}_q^d(\mu, \sigma^2)$, avec les paramètres réels $\mu = 4$ et $\sigma = 0,8$ et pour différentes valeurs de $q$. La Figure 1 présente un résumé des résultats obtenus à partir de notre simulation. D'après la Figure 1, lorsque $q$ s'approche de 1 et de 0, la distribution discrète $q$-Gaussienne présente le même histogramme que les distributions Gaussienne et Uniforme ordinaires, respectivement, comme l'indique le Théorème 1.1.

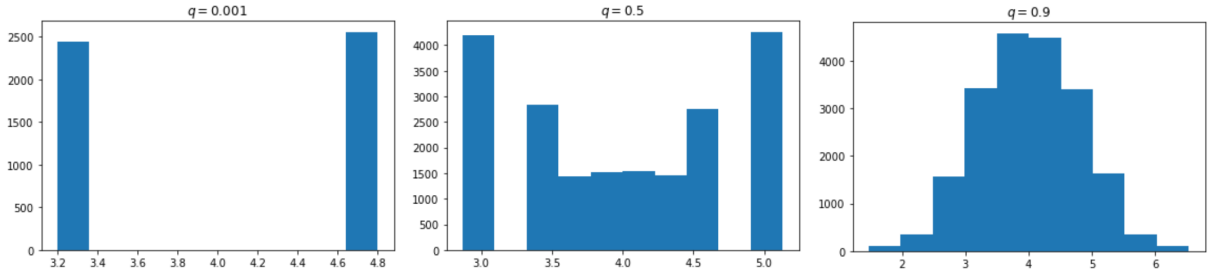



FIGURE 1 – Histogramme d'un échantillon suivant $\mathcal{N}_q^d(4, 0.8^2)$ pour différentes valeurs de $q$.

TABLE 1 – Estimation de $q$ obtenue à l'aide de la méthode des moments (MM) pour un échantillon de taille $N = 1000$.

| Modèle | $\mathcal{N}_q^d(0, 0.1^2)$ | | | $\mathcal{N}_q^d(-4, 1.2^2)$ | | | $\mathcal{N}_q^d(4, 0.8^2)$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $q$ | 0.2 | 0.5 | 0.9 | 0.2 | 0.5 | 0.9 | 0.2 | 0.5 | 0.9 |
| $\hat{q}$ | 0.1995 | 0.5002 | 0.9036 | 0.2008 | 0.50006 | 0.9022 | 0.1995 | 0.4997 | 0.902 |

Le tableau ci-dessous présente les estimations de $q$ pour trois modèles $q$-Gaussiens discrets, pour différentes valeurs de $q$ et une taille d'échantillon égale à $N = 1000$, en utilisant la méthode des moments considérée dans l'Équation (2.1). D'après le Tableau 1, le paramètre estimé $\hat{q}$ est très proche de la valeur réelle de $q$.
Afin d'évaluer la performance des deux méthodes d'estimation, à savoir la méthode des moments (MM) et la méthode du maximum de vraisemblance (MLE) présentées dans la Section 2, nous avons mené une étude de simulation à l'aide de 500 échantillons de tailles différentes, dont $N =$ 200, 500 et 1000 observations, qui ont été générés à partir des différentes distributions discrètes $q$-Gaussiennes $\mathcal{N}_q^d(\mu, \sigma^2)$. Sur la base de ces échantillons, nous avons calculé l'erreur quadratique moyenne (Root Mean Squared Error RMSE) définie comme suit

$$RMSE(\hat{\mu}, \mu) = \sqrt{MSE(\hat{\mu}, \mu)},$$

où, l'erreur quadratique moyenne (Mean Squared Error MSE) est définie par

$$MSE(\hat{\mu}, \mu) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}^i - \mu\right)^2.$$

De même, nous avons calculé (RMSE) pour l'écart-type $\sigma$, noté, $RMSE(\hat{\sigma}, \sigma)$.

Le Tableau 2 présente l'erreur quadratique moyenne relative (RMSE) de $\mu$ et de $\sigma$ calculée à l'aide des Équations (3) et (3). Comme le montre le Tableau 2, pour tous les modèles considérés, la RMSE diminue avec l'augmentation de la taille de l'échantillon, ce qui conduit à des estimations plus proches aux vrais paramètres. En outre, les deux méthodes d'estimation donnent des valeurs similaires pour chaque modèle, avec une performance légèrement meilleure pour la méthode (MM) que pour la méthode (MLE). Cependant, sauf dans les cas où $q = 0,2$, la méthode des moments fournit des résultats significativement plus précis que la méthode du maximum de vraisemblance pour l'estimation de $\sigma$.

TABLE 2 – Erreur quadratique moyenne relative (RMSE) des paramètres $\mu$ et $\sigma$.

| Modèle | $q$ | $N$ | MM | | MLE | |
|---|---|---|---|---|---|---|
| | | | RMSE($\mu$,$\hat{\mu}$) | RMSE($\sigma$,$\hat{\sigma}$) | RMSE($\mu$,$\hat{\mu}$) | RMSE($\sigma$,$\hat{\sigma}$) |
| $\mathcal{N}_q^d(0,0.1^2)$ | 0.2 | 200 | **0.00683** | **0.00176** | 0.0073 | 0.06613 |
| | | 500 | **0.00436** | **0.00112** | 0.00555 | 0.06572 |
| | | 1000 | **0.00321** | **0.00078** | 0.00396 | 0.06502 |
| | 0.5 | 200 | **0.00705** | 0.00307 | 0.00715 | **0.00216** |
| | | 500 | 0.0043 | **0.00195** | **0.00427** | 0.00212 |
| | | 1000 | **0.00308** | **0.000001** | 0.00329 | 0.00211 |
| | 0.9 | 200 | **0.00717** | 0.00471 | 0.00722 | **0.00469** |
| | | 500 | 0.00458 | 0.00308 | **0.00451** | **0.00293** |
| | | 1000 | **0.00302** | **0.00208** | 0.00321 | 0.00213 |
| $\mathcal{N}_q^d(-4,1.2^2)$ | 0.2 | 200 | **0.08305** | **0.02281** | 0.08765 | 0.08116 |
| | | 500 | **0.05604** | **0.01424** | 0.06786 | 0.08108 |
| | | 1000 | **0.03824** | **0.00916** | 0.05319 | 0.08021 |
| | 0.5 | 200 | **0.08643** | **0.03825** | 0.08861 | 0.04362 |
| | | 500 | **0.0547** | **0.02405** | 0.05526 | 0.03251 |
| | | 1000 | 0.039 | **0.01611** | **0.03868** | 0.02753 |
| | 0.9 | 200 | 0.08472 | 0.05621 | **0.08387** | **0.05402** |
| | | 500 | 0.05416 | 0.03489 | **0.05306** | **0.03393** |
| | | 1000 | 0.03752 | 0.02509 | **0.03713** | **0.02446** |
| $\mathcal{N}_q^d(4,0.8^2)$ | 0.2 | 200 | **0.05365** | **0.01485** | 0.05739 | 0.04732 |
| | | 500 | **0.03509** | **0.00858** | 0.04328 | 0.04581 |
| | | 1000 | **0.02497** | **0.00631** | 0.03441 | 0.04212 |
| | 0.5 | 200 | **0.05699** | **0.02529** | 0.05864 | 0.04515 |
| | | 500 | **0.03734** | **0.01666** | 0.03737 | 0.04166 |
| | | 1000 | **0.02436** | **0.01082** | 0.02683 | 0.03791 |
| | 0.9 | 200 | **0.05438** | **0.03682** | 0.06281 | 0.03756 |
| | | 500 | 0.03666 | **0.02362** | **0.03641** | 0.02482 |
| | | 1000 | **0.0232** | **0.016** | 0.02429 | 0.01613 |

Les paramètres estimés à l'aide de l'algorithme (E-M) d'un mélange de trois distributions $q$-Gaussienne discrètes pour une valeur fixée de $q$ sont présentés dans le tableau suivant. Le Tableau

TABLE 3 – Paramètres estimés d'un mélange de trois $q$-Gaussiennes discrètes.

| Paramètres | Vrais paramètres | Paramètres estimés |
|---|---|---|
| $\pi$ | $[0.37, 0.34, 0.29]$ | $[0.373, 0.336, 0.291]$ |
| $\mu$ | $[0, -4, 4]$ | $[-0.051, -3.974, 3.981]$ |
| $\sigma$ | $[0.1, 1.2, 0.8]$ | $[0.946, 1.198, 0.787]$ |

3, montre que les valeurs estimées des paramètres sont proches des valeurs réelles, ce qui indique que l'algorithme (E-M) est robuste et il fournit des estimations précises pour les paramètres d'un mélange fini de $q$-Gaussiennes discrètes.

# 4 conclusion

Dans ce travail, nous avons introduit une nouvelle distribution discrète, appelée distribution discrète $q$-Gaussienne, notée par $\mathscr{N}_q^d(0,1)$. Nous avons généralisé cette distribution à $\mathscr{N}_q^d(\mu,\sigma^2)$. Par la suite, nous avons estimé le paramètre $q$ à l'aide de la méthode des moments, et les nouveaux paramètres $\mu$ et $\sigma^2$ à la fois à l'aide de la méthode des moments et la méthode du maximum de vraisemblance. Ce qui implique une estimation des paramètre du modèle $q$-Gaussien $\mathscr{N}_q(\mu,\sigma^2)$. Ensuite, nous avons examiné un mélange fini de $q$-Gaussiennes discrètes tout en estimant ses paramètres en appliquant l'algorithme (E-M). Enfin, nous avons réalisé une étude de simulation afin d'évaluer la performances du modèle et des estimateurs proposés.

# Bibliographie


Ben Mrad, O., Masmoudi, A., et Slaoui, Y. (2023), *Some properties of q-Gaussian distributions*, Communications in Statistics - Theory and Methods.

Ben Mrad, O., Masmoudi, A., et Slaoui, Y. (2024), *The discrete q-Gaussian distribution $\mathscr{N}_q(\mu,\sigma^2)$ : Properties and parameters estimation*, Physics Letters A.

Díaz, R. et Pariguan, E. (2009), *On the Gaussian q-distribution*, Journal of Mathematical Analysis and Applications, 357, pp. 1-9.

Jackson, F. H. (1910), *q-Difference Equations*, American Journal of Mathematics, 32, pp. 305-314.

Kac, V. et Cheung, P. (2002), *Quantum calculus*, Universitext. New York : Springer-Verlag.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977), *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the royal statistical society : series B (methodological), 39, pp. 1-22.

Díaz, R., Ortiz, C. et Pariguan, E. (2010), *On the k-gamma q-distribution*, Central European Journal of Mathematics, 8, pp. 448-458.

Díaz, R. et Pariguan, E. (2007), *On hypergeometric functions and Pochhammer k-symbol*, Revista Matemática de la Universidad del Zulia. Divulgaciones Matemáticas, 15, pp. 179-192.

Díaz, R. et Teruel, C. (2005), *q,k-generalized gamma and beta functions*, Journal of Nonlinear Mathematical Physics, 12, pp. 118-134.

Fitouhi, A. and Bettaibi, N. (2007), *Applications of the Mellin transform in quantum calculus*, Journal of Mathematical Analysis and Applications, 328, pp. 518-534.

Yamano, T. (2002), *Some properties of q-logarithm and q-exponential functions in Tsallis statistics*, Physica A : Statistical Mechanics and its Applications, 305, pp. 486-496.

# Inférence Post-Clustering

Nicolas Enjalbert-Courrech [1,a] & Cathy Maugis-Rabusseau [1,b] & Pierre Neuvial [1,c]

[1] *Institut de Mathématiques de Toulouse; UMR5219 Université de Toulouse; CNRS,*
[a] *UPS, F-31062 Toulouse Cedex 9, France France*
*nicolas.enjalbert-courrech@math.univ-toulouse.fr*
[b] *INSA, F-31077 Toulouse, France*
*cathy.maugis@insa-toulouse.fr*
[c] *UPS, F-31062 Toulouse Cedex 9, France France*
*pierre.neuvial@math.univ-toulouse.fr*

**Résumé.** On s'intéresse au problème de "double-dipping" c'est-à-dire à l'utilisation du même jeu de données pour faire d'abord un clustering des observations puis un test statistique dont l'hypothèse nulle dépend des classes obtenues à l'étape précédente. Dans un premier temps, nous effectuons un état de l'art des nombreuses approches récemment proposées pour ce problème, en les regroupant en deux catégories : les méthodes de partitionnement de l'information et les approches conditionnelles. Ensuite, nous proposons une comparaison numérique afin d'évaluer leur performance en termes de contrôle du risque de première espèce, de puissance statistique, et de temps de calcul.

**Mots-clés.** Tests d'hypothèse, clustering, double-dipping, test post-sélection.

**Abstract.** This work tackles the problem of "double-dipping," which refers to the use of the same dataset to first perform clustering of observations and then conduct a statistical test, where the null hypothesis depends on the clusters obtained in the previous step. Initially, we conduct a review of the numerous approaches recently proposed for this problem, grouping them into two categories: information partitioning methods and conditional approaches. Then, we propose a numerical comparison to evaluate their performance in terms of controlling the Type I error rate, statistical power, and computation time.

**Keywords.** Hypothesis testing, clustering, double-dipping, selective inference.

# 1 Introduction

Dans ce travail, nous nous intéressons au problème de "double-dipping" c'est-à-dire à l'utilisation des mêmes données pour faire 1) un clustering des individus puis 2) un test basé sur les classes obtenues. Nous nous plaçons dans un cadre Gaussien où pour chaque individu $i \in \{1, \dots, n\}$, $X_i \sim \mathcal{N}_p(\mu_i, \Sigma)$ et les $X_i$ sont indépendants. On note par la suite $\mathbf{X} = (X_i)_{i=1,\dots,n}$ la matrice de taille $n \times p$ regroupant les vecteurs $X_i$. Soit $C(\mathbf{X}) = \{C_1(\mathbf{X}), ..., C_K(\mathbf{X})\}$ le clustering en $K$ classes obtenues

par une méthode de clustering $C$ sur $\mathbf{X}$. Dans ce travail, on s'intéresse à la question de tester s'il y a une différence entre deux classes $C_k(\mathbf{X})$ et $C_{k'}(\mathbf{X})$. On considère donc l'hypothèse nulle :

$$\mathcal{H}_0^\eta : \eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^T \boldsymbol{\mu} = 0, \tag{1}$$

avec $\boldsymbol{\mu} = (\mu_i)_{i=1,\dots,n} \in \mathbb{R}^{n\times p}$ et le vecteur de contraste $\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})) \in \mathbb{R}^n$. Par exemple, pour le test de comparaison des moyennes des classes $C_k$ et $C_{k'}$, le vecteur de contraste s'écrit pour chaque individu $i$,

$$\eta_i(C_k, C_{k'}) = \left( \frac{\mathbb{1}_{i\in C_k}}{|C_k|} - \frac{\mathbb{1}_{i\in C_{k'}}}{|C_{k'}|} \right) \tag{2}$$

avec $|C_k|$ le cardinal de la classe $C_k$.

Lorsque le vecteur de contraste $\eta$ est fixé a priori et ne dépend pas du jeu de données observé, l'hypothèse nulle n'est pas aléatoire et les tests classiques de comparaison de moyennes entre deux classes peuvent être appliqués. Par exemple, la $p$-valeur $p(\mathbf{x})$ associée au test de Hotelling est donnée par

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0^\eta}\left( ||\eta^T \mathbf{X}||_2 \geq ||\eta^T \mathbf{x}||_2 \right). \tag{3}$$

Dans le cadre étudié ici, le vecteur $\eta = \eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))$ dépend des données, et la procédure de test "naïve" associée à la $p$-valeur dans (3) ne contrôle plus le risque de première espèce [Gao et al., 2022].

Les méthodes récemment proposées pour répondre à cette problématique de test post-clustering peuvent être classées en deux catégories : les méthodes de partitionnement de l'information et les approches conditionnelles. L'objectif de ce travail est de faire un état de l'art de ces méthodes et de comparer ces méthodes par des simulations numériques.

# 2 Partitionnement de l'information

L'idée de la première catégorie des méthodes étudiées est de reprendre le principe de partitionnement des données pour les méthodes d'apprentissage supervisé proposé par Cox [1975]. Le but est d'obtenir deux sous-échantillons indépendants, l'un pour construire/ajuster le modèle et l'autre pour la procédure de test. Dans le cas de l'inférence post-clustering, Zhang et al. [2019] ont développé une procédure de test qui utilise le *data splitting* mais cette procédure ne contrôle pas le risque de première espèce. En effet, reporter l'information du clustering (construit à partir du premier sous-échantillon) afin de labéliser les individus du deuxième sous-échantillon avant le test, transpose une information provenant du premier échantillon sur le deuxième. Comme le montre Gao et al. [2022], l'hypothèse testée reste aléatoire du fait du lien entre le vecteur de contraste et les données utilisées pour faire le test.

Afin de contrebalancer le transfert d'information pour la labélisation du jeu de test, Leiner et al. [2023] et Neufeld et al. [2023a] proposent de nouvelles méthodes de séparation des données. Au lieu de partitionner le jeu de données $\mathbf{X}$ en deux sous-ensembles d'observations de taille $n_1$ et $n_2$

avec $n_1 + n_2 = n$, ces méthodes partitionnent l'information pour chaque observation, construisant ainsi deux jeux de données indépendants ou conditionnellement indépendants $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ de taille $n$. Le clustering appliqué sur $\mathbf{X}^{(1)}$ permet d'obtenir la classification des $n$ individus, et d'effectuer la procédure de test sur $\mathbf{X}^{(2)}$ indépendant du vecteur de contraste.

Leiner et al. [2023] proposent une procédure appelée *data fission* qui vise à créer deux nouveaux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ tels que $\mathbf{X} = h(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, où les lois de $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}|\mathbf{X}^{(1)}$ sont connues et $h()$ est une fonction connue. Pour la loi gaussienne multivariée $X_i \sim \mathcal{N}(\mu_i, \Sigma)$, la procédure définit $Z_i \sim \mathcal{N}(0, \Sigma)$ indépendante de $X_i$, et $X_i^{(1)} = X_i + \tau Z_i \sim \mathcal{N}(\mu_i, (1+\tau^2)\Sigma)$ et $X_i^{(2)} = X_i - \tau^{-1} Z_i \sim \mathcal{N}(\mu_i, (1+\tau^{-2})\Sigma)$. Le paramètre de fission $\tau > 0$ permet de définir le partage d'information entre $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$.

Neufeld et al. [2023a] proposent la procédure de *data thinning* qui, pour certaines lois, permet de décomposer $\mathbf{X}$ en deux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ indépendants tels que $\mathbf{X}^{(1)} + \mathbf{X}^{(2)} = \mathbf{X}$. Pour une loi gaussienne multivariée, la procédure propose de générer $X_i^{(1)}|X_i = x_i \sim \mathcal{N}_p(\epsilon x_i, \epsilon(1-\epsilon)\Sigma)$, où $\epsilon \in [0, 1]$ est un paramètre de partage de l'information, puis $X_i^{(2)} = X_i - X_i^{(1)}$. Ainsi la procédure donne deux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ de lois connues et indépendants. Le clustering peut ainsi être obtenu à partir de $\mathbf{X}^{(1)}$ tel que $C_k, C_{k'} \in C(\mathbf{X}^{(1)})$, et $\eta(C_k, C_{k'})$ est considéré comme fixé pour la procédure de test sur $\mathbf{X}^{(2)}$. La loi sous l'hypothèse nulle de la $p$-valeur

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0^\eta}\left(||\eta(\mathbf{x}^{(1)})^T \mathbf{X}^{(2)}||_2 \geq ||\eta(\mathbf{x}^{(1)})^T \mathbf{x}^{(2)}||_2\right) \tag{4}$$

est ainsi connue.

Néanmoins, ces deux méthodes demandent de connaître la vraie matrice de covariance $\Sigma$. De plus, le paramètre $\epsilon$ (resp. $\tau$), qui pilote le partage d'information, a besoin d'être calibré dans la procédure de *data thinning* (resp. *data fission*).

# 3 Approche conditionnelle

## 3.1 Définition d'un test conditionnel

Une seconde famille de méthodes permettant de résoudre ce problème d'inférence post-clustering consiste à prendre en compte explicitement l'action de clustering dans le calcul de la $p$-valeur. Celle-ci est inspirée de la littérature récente sur l'inférence post-sélection de modèle [Fithian et al., 2014], où il s'agit de prendre en compte une étape de sélection de variables. Par exemple pour le test d'Hotelling considéré dans (3), on souhaite conditionner par l'évènement $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in C(\mathbf{X})\}$ et définir une $p$-valeur conditionnelle comme :

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_0^\eta}\Big(||\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^T \mathbf{X}||_2 \geq ||\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^T \mathbf{x}||_2 \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in C(\mathbf{X})\Big) \\
&= \mathbb{P}_{\mathcal{H}_0^\eta}\Big(||\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^T \mathbf{X}||_2 \geq ||\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^T \mathbf{x}||_2 \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in C(\mathbf{X})\Big).
\end{aligned} \tag{5}$$

Ce conditionnement dans (5) permet de fixer le vecteur de contraste, qui ne dépend que du résultat du clustering. Néanmoins cette $p$-valeur n'étant pas accessible, il faut choisir un conditionnement plus fort donnant accès à la loi de la statistique de test sous l'hypothèse nulle.

Gao et al. [2022] s'intéressent au problème de comparaison de moyenne entre classes (voir équation (2)) sous l'hypothèse $\Sigma = \sigma^2 I_p$. Afin de pouvoir expliciter la $p$-valeur, ils sur-conditionnent par rapport à (5) en s'appuyant sur la décomposition $\mathbf{X} = \pi_\eta^\perp \mathbf{X} + \frac{\|\eta^T \mathbf{X}\|_2}{\|\eta\|_2^2}\, \eta\, dir(\eta^T \mathbf{X})$ où $\pi_\eta^\perp$ est la projection sur l'orthogonal de la droite engendrée par le vecteur de contraste $\eta$. En fixant $\pi_\eta^\perp \mathbf{X}$ et $dir(\eta^T \mathbf{X})$, l'aspect aléatoire dans la décomposition de $\mathbf{X}$ est uniquement porté par la statistique de test $\|\eta^T \mathbf{X}\|_2$. La $p$-valeur obtenue par ce sur-conditionnement peut alors s'écrire :

$$p(\mathbf{x}, \{C_k, C_{k'}\}) = 1 - \mathbb{F}(\|\eta^T \mathbf{x}\|_2; \sigma\|\eta\|_2, S(\mathbf{x}, \{C_k; C_{k'}\})) \tag{6}$$

où $\mathbb{F}(t; c, S)$ désigne la fonction de répartition de la loi $c.\chi_p$ tronquée à un ensemble $S$, et

$$S(\mathbf{x}, \{C_k, C_{k'}\}) = \{\phi \geq 0 : C_k, C_{k'} \in C(\tilde{\mathbf{x}}(\phi))\} \tag{7}$$

est l'ensemble des $\phi > 0$ tels que les classes $C_k$ et $C_{k'}$ sont préservées par la procédure de clustering appliquée aux données perturbées $\tilde{\mathbf{x}}(\phi) = \pi_\eta^\perp \mathbf{x} + \frac{\phi}{\|\eta\|_2^2}\, \eta\, dir(\eta^T \mathbf{x})$.

## 3.2 Mise en pratique de la méthode

Gao et al. [2022] obtiennent une caractérisation explicite de $S$ dans le cas de la classification ascendante hiérarchique pour certaines mesures d'agrégation, et arrivent ainsi à un calcul explicite de la $p$-valeur associée. Chen and Witten [2023] ont étendu cette procédure dans le cadre de la classification par $K$-means. Pour réussir à expliciter l'ensemble $S$, ils s'appuient sur les propriétés des $K$-means et en conditionnant par un événement plus fort imposant le maintien de la partition des individus à chaque itération de l'algorithme des $K$-means, appliqué aux données perturbées. Néanmoins, ce sur-conditionnement risque de faire perdre de la puissance au test statistique final. Lorsque cet ensemble $S$ ne peut être explicité, une procédure de Monte-Carlo par Importance Sampling est utilisée pour approcher la $p$-valeur. Bien que ces résultats aient été obtenus spécifiquement dans le cas de la comparaison de deux classes (voir équation (2)), nous avons montré qu'ils se généralisent sans difficulté à n'importe quel vecteur de contraste ne dépendant que des classes $C_k(\mathbf{X})$ et $C_{k'}(\mathbf{X})$.

## 3.3 Extensions pour relaxer la variance sphérique connue

Dans un premier temps, Gao et al. [2022] et Chen and Witten [2023] établissent leurs résultats pour une matrice de covariance sphérique connue $\Sigma = \sigma^2 I_p$. Ils proposent ensuite d'étendre les résultats au cas d'une variance générale connue en transformant les données afin de les rendre sphérique. Si la matrice de covariance n'est pas connue, Gao et al. [2022] et Chen and Witten [2023] ont étudié théoriquement l'impact de l'estimation de la variance dans le cas sphérique sur le contrôle du risque de première espèce. Estimer la variance en ignorant la structure de classe conduit à une surestimation de la variance. Gao et al. [2022] ont montré qu'une telle surestimation maintient le contrôle du risque de première espèce, au prix d'une perte de puissance. González-Delgado et al. [2023] ont proposé une généralisation de la définition du test conditionnel de Gao et al. [2022], et ont étendu le résultat de surestimation ci-dessus à une covariance $\Sigma$ quelconque, et pour toute structure de dépendance entre individus.

Afin de contourner le problème difficile d'estimation de la variance, Yun and Foygel Barber [2023] proposent un test conditionnel dans le cas d'une variance sphérique inconnue. Pour aborder ce problème, ils considèrent une hypothèse nulle plus forte, demandant l'égalité de toutes les vraies moyennes $\mu_i$ de tous les individus $i \in C_k(\mathbf{X}) \cup C_{k'}(\mathbf{X})$. Une nouvelle statistique de test est proposée, qui prend en compte la dispersion intra-classe des individus :

$$R(\mathbf{X}) = (|C_k(\mathbf{X})| + |C_{k'}(\mathbf{X})| - 2)\,\frac{\|\mathcal{P}_0\mathbf{X}\|_F^2}{\|\mathcal{P}_1\mathbf{X}\|_F^2} \tag{8}$$

où $\mathcal{P}_0\mathbf{X}$ capture la différence des moyennes (comme dans Gao et al. [2022]) et $\mathcal{P}_1\mathbf{X}$ capture l'inertie intra-classe entre les deux classes considérées. Dans le même esprit que Gao et al. [2022], $\mathbf{X}$ est décomposé en fonction de $\mathcal{P}_0\mathbf{X}$, $\mathcal{P}_1\mathbf{X}$ et $(I_n - \mathcal{P}_0 - \mathcal{P}_1)\mathbf{X}$. La $p$-valeur est alors sur-conditionnée en fixant les valeurs des éléments suivants :

$$\|\mathcal{P}_0\mathbf{X}\|_F^2 + \|\mathcal{P}_1\mathbf{X}\|_F^2, \frac{\mathcal{P}_0\mathbf{X}}{\|\mathcal{P}_0\mathbf{X}\|_F}, \frac{\mathcal{P}_1\mathbf{X}}{\|\mathcal{P}_1\mathbf{X}\|_F}, (I_n - \mathcal{P}_0 - \mathcal{P}_1)\mathbf{X}. \tag{9}$$

Ainsi cette $p$-valeur sur-conditionnée peut être calculée à partir d'un quantile d'une loi de Fisher tronquée à un nouvel ensemble $S$ préservant le clustering sur des données perturbées. Dans le cas d'un clustering à 2 classes, pour les $K$-means et Classifications Ascendantes Hiérarchiques , le nouvel ensemble $S$ est inclus dans l'ensemble $S$ explicite, proposé par Chen and Witten [2023] et Gao et al. [2022] respectivement. Dans le cas d'un clustering à plus de 2 classes ou une autre méthode de clustering, la $p$-valeur est estimée par Importance Sampling.

# 4 Comparaison numérique des performances des méthodes

Un grand nombre de publications récentes sont apparues sur ce sujet [Gao et al., 2022, Chen and Witten, 2023, González-Delgado et al., 2023, Yun and Foygel Barber, 2023, Leiner et al., 2023, Neufeld et al., 2023a, Dharamshi et al., 2023], complété par des publications d'application des méthodes à des problématiques biologiques [Neufeld et al., 2024, 2023b] ou des publications adaptant les méthodes à des tests de comparaison de classe par variable [Hivert et al., 2024, Chen and Gao, 2023].

Dans ce contexte, nous avons souhaiter proposer une comparaison quantitative des différentes méthodes au travers de simulations numériques. Dans un premier temps, l'étude que nous avons menée permet de vérifier si ces méthodes contrôlent bien le risque de première espèce. Une fois les méthodes défaillantes mises de côté, l'étude propose une analyse de la puissance statistique des méthodes toujours en compétition. Il s'agit en particulier de mesurer l'impact sur la puissance statistique et sur le temps de calcul d'une méthode d'inférence conditionnelle avec une expression explicite de la $p$-valeur, comparée à une estimation par Importance Sampling.

L'objectif de telles expériences numériques est de 1) comprendre et trouver le meilleur compromis entre la puissance statistique et le temps de calcul des méthodes étudiées et 2) prendre du recul sur la possibilité d'étendre le calcul explicite à d'autres méthodes de clustering. En particulier, une perspective naturelle de ce travail est l'étude du clustering par mélanges gaussiens. En effet, cette méthode de clustering permet d'estimer la matrice de covariance de chaque classes, et d'obtenir

une probabilité d'appartenance de chaque individu à une classe. Il serait intéressant d'exploiter ces deux informations dans le cadre de l'inférence post clustering.

# References


Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11, 2022.

David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2): 441–444, 1975.

Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, 9(4):383–392, 2019.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *Journal of the American Statistical Association*, pages 1–12, 2023.

Anna Neufeld, Ameer Dharamshi, Lucy L Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*, 2023a.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Yiqun T Chen and Daniela M Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

Javier González-Delgado, Juan Cortés, and Pierre Neuvial. Post-clustering inference under dependency. *arXiv preprint arXiv:2310.11822*, 2023.

Young-Joo Yun and Rina Foygel Barber. Selective inference for clustering with unknown variance. *Electronic Journal of Statistics*, 17(2):1923–1946, 2023.

Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L Gao, Daniela Witten, and Jacob Bien. Generalized data thinning using sufficient statistics. *arXiv preprint arXiv:2303.12931*, 2023.

Anna Neufeld, Lucy L Gao, Joshua Popp, Alexis Battle, and Daniela Witten. Inference after latent variable estimation for single-cell rna sequencing data. *Biostatistics*, 25(1):270–287, 2024.

Anna Neufeld, Joshua Popp, Lucy L Gao, Alexis Battle, and Daniela Witten. Negative binomial count splitting for single-cell rna sequencing data. *arXiv preprint arXiv:2307.12985*, 2023b.

Benjamin Hivert, Denis Agniel, Rodolphe Thiébaut, and Boris P Hejblum. Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, page 107916, 2024.

Yiqun T Chen and Lucy L Gao. Testing for a difference in means of a single feature after clustering. *arXiv preprint arXiv:2311.16375*, 2023.

# Fission de données pour l'inférence post-classification : de la théorie à la pratique

Benjamin Hivert[1,2], Denis Agniel[3], Rodolphe Thiébaut[1,2,4] & Boris Hejblum[1,2]

[1] *Univ. Bordeaux, INSERM, INRIA, SISTM team, Bordeaux Population Health, U1219, F-33000 Bordeaux, France*
[2] *Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France*
[3] *Rand Corporation, Santa Monica, CA 90401, USA*
[4] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*

*benjamin.hivert@u-bordeaux.fr, dagniel@rand.org, rodolphe.thiebaut@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr*

**Résumé.** Dans divers domaines, tels qu'en génomique, la classification non supervisée pose des défis en raison de son utilisation pour formuler des hypothèses testées sur les mêmes ensembles de données. Cette pratique, appelée inférence post-classification, compromet les propriétés statistiques des tests, en particulier le contrôle de l'erreur de Type I. La fission de données (Leiner et al., 2023) permet d'obtenir deux jeux de données indépendants à partir d'un seul échantillon, en bruitant l'information contenue dans chaque observation en deux parties de manière précise. Ainsi, chaque partie est utilisable indépendamment (pour la classification non supervisée d'une part, et pour les tests d'hypothèses qui suivent d'autre part) sans impacter les propriétés habituelles des méthodes statistiques. Cependant, ses limitations, notamment en termes d'hypothèses distributionnelles et d'adaptabilité à des situations avec de véritables classes inconnues, restreignent son utilisation. L'application rigoureuse de la fission de données pour l'inférence post-classification exige une connaissance préalable des véritables classes et des variances intra-classes associées. Or ces informations sont inconnues en pratique et doivent alors être estimées. Nous démontrons que l'indépendance promise par la fission de données n'est garantie qu'à condition de posséder des estimations non-biaisés des variances, et que par conséquent, elle ne peut théoriquement assurer le contrôle de l'erreur de Type I des tests associés. Nous proposons une nouvelle approche consistant à modéliser chaque observation comme une réalisation d'un processus distinct, avec des paramètres individuels, que nous estimons alors de manière non-paramétrique. Les performances de cette nouvelle approche ont été évaluées au moyen de simulations numérique, révélant l'absolue nécessité d'une très bonne séparation entre les classes afin de garantir des estimations non-biaisées des variances locales, et donc le contrôle effectif de l'erreur de Type I associée. En conclusion, bien que la fission de données ait été initialement envisagée comme une solution aux problèmes d'inférence post-classification, sa mise en pratique est rendue extrêmement délicate par le lien entre l'estimation de la structure des vraies classes et celle de leur variance. Le bon comportement de cette approche pour l'inférence post-classification nécessite indirectement de connaître les vraies classes – cependant inconnues – que l'on cherche aussi à estimer. Notre nouvelle approche de modélisation résout cette difficulté dans certains cas favorables, mais elle souffre des difficultés inhérentes à l'estimation non paramétrique de la variance locale.

**Mots-clés.** Inférence post-classification, Fission de données, Estimation non-paramétrique, Variance locale

**Abstract.** In various fields, such as genomics, clustering poses challenges due to its use in formulating hypotheses tested on the same datasets. This practice, known as post-clustering inference, compromises the statistical properties of tests, particularly the Type I error control. To address this, data fission provides an innovative approach by decomposing the information contained in each observation into two parts, independently usable for clustering and subsequent hypothesis testing. However, its limitations, especially in terms of distributional assumptions and adaptability to situations with true unknown classes, restrict its application. In this context, the rigorous application of data fission requires prior knowledge of the true classes and associated intra-class variances. However, in real applications, this information is unknown and must be estimated from the data. We demonstrate that the independence guaranteed by data fission is only valid for unbiased estimators of variance. Therefore, it cannot theoretically ensure control of the Type I error of associated tests due to the complexity of unbiased estimation of unknwow intra-class variances required for its application. Facing these challenges, we propose an alternative approach, modeling each observation as a realization of its own distribution with specific parameters estimated in a non-parametric manner. The performance of this new approach was evaluated through simulations, revealing the need for a clear separation between classes to ensure unbiased estimations and thus effective control of Type I error. In conclusion, although data fission was initially proposed as a solution to post-clustering inference problems, its applicability is compromised by the need to know the true classes. Indeed, the proper behavior of this approach for post-clustering inference depends directly on the true unknown classes to be estimated. The new modeling approach we propose allows overcoming this need for prior knowledge but presents challenges related to non-parametric variance estimation.

**Keywords.** Post-clustering inference, Data fission, Non-parametric estimation, Local variance

# 1 Introduction

Lors d'analyses exploratoires, on applique couramment des méthodes de classification non supervisée pour identifier la structure des données en regroupant les observations en sous-groupes (appelés « classes ») homogènes et séparés. Ces classes peuvent ensuite être utilisées pour générer des hypothèses, qui seront testées par des tests statistiques. Par exemple, il est courant en transcriptomique d'effectuer des tests univariés pour identifier les variables (i.e. les gènes) dont l'expression est significativement associée à une classe, dans le but de décrire et d'interpréter cette dernière. Cependant, étant donné que les classes sont obtenues à partir des mêmes données que celles utilisées pour les tests, ces analyses en deux étapes ne respectent plus le cadre théorique des tests statistiques, où les hypothèses de tests doivent être posées avant les analyses et ne peuvent dépendre des données. Ce problème connu sous le nom de « double-dipping »(Kriegeskorte et al., 2009) en anglais, compromet les bonnes

propriétés statistiques des tests utilisés. En particulier dans le cas de tests post-classification, on observe alors un mauvais contrôle de l'erreur de Type I (Gao et al., 2022). Récemment, des tests d'hypothèses spécialement conçus pour résoudre les problèmes d'inférence post-classification ont été proposés (Zhang et al., 2019; Chen and Gao, 2023; Hivert et al., 2024). Ils assurent un contrôle effectif de l'erreur de Type I dans ce contexte de double utilisation des données, en se basant notamment sur les concepts d'inférences sélectives. Cependant, leur application pratique est entravée par leur temps de calcul conséquent, et leur restriction à des méthodes de classification non-supervisée particulières, ou la distribution et la dimension des données.

Leiner et al. (2023) ont récemment introduit une approche attractive permettant l'utilisation répétée d'observations issues d'un même échantillon : la fission de données. Comparable aux divisions classiques en échantillons respectivement d'apprentissage et de test fréquemment utilisées en apprentissage automatique, cette méthode vise à décomposer l'information contenue dans chaque observation en deux parties indépendantes. Il est possible d'utiliser la première pour construire les classes, et la seconde pour tester des différences entre elles. Cependant, cette méthode repose sur des hypothèses distributionnelles fortes, et seules deux distributions (la distribution de Poisson et la distribution gaussienne) bénéficient d'une procédure de fission directement exploitable dans le contexte de l'inférence post-classification. Plus récemment, Neufeld et al. (2023) ont élargi la versatilité des distributions pouvant être décomposées en généralisant la fission de données grâce au « data thinnig ». Bien que théoriquement la fission de données et sa généralisation aient été proposées pour répondre aux défis du double-dipping dans le cadre de la classification non-supervisée, nous démontrons ici que ces méthodes sont extrêmement difficiles à appliquer en pratique. Elles reposent sur des hypothèses distributionnelles exigeant une absence de classe, et adapter ces méthodes aux situations impliquant de véritables classes inconnues requiert l'estimation d'hyperparamètres spécifiques à chacune de ces classes inconnues. Nous mettons alors en lumière l'impact d'une estimation biaisée de ces hyperparamètres sur l'erreur de Type I des tests suivant la classification non-supervisée, malgré l'utilisation de la fission de données. Nous proposons également une stratégie d'estimation non-paramétrique de la variance locale pour le cas gaussien, offrant ainsi la possibilité de s'affranchir de la nécessité de connaître les vraies classes dans certains cas favorables.

# 2 Méthode

## 2.1 Décompositions en variables aléatoires indépendantes

Soit $X$ une variable aléatoire. La fission de données proposée par Leiner et al. (2023) a pour objectif de décomposer $X$ en deux nouvelles variables aléatoires $X^{(1)}$ et $X^{(2)}$ grâce à l'ajout de deux bruits paramétriques précisément reliés. Ces deux transformations contiennent chacune de l'information sur $X$, mais incomplète. De plus, la répartition de la quantité d'information issue de $X$ entre $X^{(1)}$ ou $X^{(2)}$ dépend d'un paramètre $\tau$. Formellement, la fission de données décompose $X$ en deux parties $X^{(1)}$ et $X^{(2)}$ telles que soit $\boldsymbol{\mathcal{P}_1}$) : $X^{(1)}$ et $X^{(2)}$ sont

indépendantes de distributions connues ; soit $\boldsymbol{\mathcal{P}_2}$) : $X^{(1)}$ a une distribution (marginale) connue et $X^{(2)}|X^{(1)}$ a une distribution (conditionnelle) connue soient vérifiées. Dans notre cadre de l'inférence post-classification, la propriété $\boldsymbol{\mathcal{P}_1}$ est requise pour garantir l'indépendance entre la classification servant à construire les classes (les hypothèses de tests) sur les réalisations de $X^{(1)}$ et les tests statistiques entre ces classes appliqués sur les réalisations de $X^{(2)}$. Seules la loi Poisson et la loi normale admettent une décomposition vérifiant $\boldsymbol{\mathcal{P}_1}$. La méthode du *data thinning* (Neufeld et al., 2023) généralise la fission de données en proposant une méthode de décomposition garantissant $\boldsymbol{\mathcal{P}_1}$ pour une famille de distributions de probabilité plus large (incluant notamment la loi Poisson, la loi normale et la loi binomiale négative). Nous allons ici d'avantage nous intéresser à la fission de données dans le cas gaussien, mais les résultats présentés plus bas sont généralisables au *data thinning* d'autres distributions de probabilité.

Soit $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ une variable aléatoire gaussienne à $p$ dimensions. Alors, la fission de $\boldsymbol{X}$ est donnée, pour $\tau \in ]0, +\infty)$, par :

$$\boldsymbol{X}^{(1)} = \boldsymbol{X} + \tau \boldsymbol{Z} \qquad \text{et} \qquad \boldsymbol{X}^{(2)} = \boldsymbol{X} - \frac{1}{\tau}\boldsymbol{Z} \qquad \text{pour } \boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{1}$$

$\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ ainsi construites respectent la propriété $\boldsymbol{\mathcal{P}_1}$ (Leiner et al., 2023). Il en découle que $\boldsymbol{X}^{(1)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, (1+\tau^2)\boldsymbol{\Sigma}\right)$ et $\boldsymbol{X}^{(2)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, (1+\frac{1}{\tau^2})\boldsymbol{\Sigma}\right)$. Ainsi, la fission de données dans le cas gaussien se traduit par la construction de deux nouvelles variables aléatoires, dont la variance est augmentée par rapport à la variable aléatoire originale $\boldsymbol{X}$.

## 2.2 Classes & variances inconnues : des limitations pratiques de la fission de données

La décomposition formulée dans l'équation (1) révèle deux limitations majeures de la fission de données, mettant en lumière la complexité de la mise en oeuvre pratique de ces approches. Tout d'abord, cette procédure de fission de données est uniquement applicable aux réalisations d'une variable aléatoire gaussienne. Cependant, cette approche de modélisation se heurte à la réalité de la présence de vraies classes inconnues dans les données. En effet, ces données sont plus généralement modélisées comme des réalisations de mélanges de gaussiennes, où chaque composante (elle-même gaussienne) représente sa propre classe. Ainsi, la fission de données, telle qu'elle a été initialement introduite, ne peut être appliquée que dans des situations où l'on présume une unique classe homogène. Dans un contexte de mélange gaussien, elle ne peut uniquement être appliquée qu'à chacune des composantes du mélange de manière indépendante. Or cela n'est faisable que si les composantes du mélange, et donc les classes, sont connues. C'est cette condition préalable de la connaissance des classes qui représente une limitation significative en pratique : la classification non-supervisée visant à identifier ces classes inconnues demeure l'une des motivations principales de l'application de la fission de données. De plus, pour appliquer la fission de données, il est nécessaire de connaître la matrice de covariance des observations $\boldsymbol{\Sigma}$. Étant inconnue, il est possible de l'estimer par la matrice de covariance empirique $\hat{\boldsymbol{\Sigma}}$. Cependant, pour appliquer la fission de données à chaque composante d'un mélange, il est alors nécessaire de considérer les matrices de covariances intra-composantes, qui sont impossibles à estimer sans la connaître la vraie structure en

classes. De surcoît, l'ensemble des propriétés théoriques de la fission de données nécessite de connaître la vraie matrice de variance $\boldsymbol{\Sigma}$, et en particulier pour l'indépendance entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$. On peut montrer que considérer une estimation de $\boldsymbol{\Sigma}$ par $\hat{\boldsymbol{\Sigma}}$ résulte en une covariance entre ces deux nouvelles variables aléatoires donnée par : $\mathrm{Cov}\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\right) = \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}$.

### 2.3 Modélisation individualisée pour une fission de données généralisée & Estimation non-paramétrique de la variance

Afin de généraliser la fission de données quelque soit le nombre de classe dans les données, nous proposons de modéliser chaque observation comme étant une réalisation de sa propre distribution avec ses propres paramètres individuels, c'est-à-dire :

$$\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{2}$$

Cette approche consiste donc à inclure des moyennes et des variances qui ne sont pas spécifiques aux classes (inconnues), mais plutôt spécifiques à chaque observation. La connaissance de la structure du mélange n'est alors plus nécessaire pour l'estimation de la variance. Seule la proximité entre individus importe, puisque cette modélisation suppose que deux individus provenant d'une même composante présentent des valeurs similaires pour ces paramètres.

Cependant, l'estimation de la variance $\boldsymbol{\Sigma}_i$ devient encore plus délicate. Les estimateurs classiques tels que la variance empirique ne sont pas adaptés. Nous proposons d'estimer ces variance de façon non-paramétrique en pondérant les observations dans leurs calculs à l'aide d'un noyau configuré de manière à ce qu'un individu proche de l'individu $i$ (et donc vraisemblablement issu de la même composante du mélange) ait un poids important mais que les individus très éloignés de l'individu $i$ se voient attribuer un poids insignifiant.

## 3 Résultats

### 3.1 Impact des Biais dans l'estimation de la variance sur la classification non-supervisée et l'erreur de Type I

Tout d'abord, il est crucial de noter qu'un biais dans l'estimation de la variance induit directement une covariance non-nulle entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$, et donc une perte de l'indépendance. Le contrôle de l'erreur de Type I pour les tests sur $\boldsymbol{X}^{(2)}$ entre les classes estimées sur $\boldsymbol{X}^{(1)}$ n'est plus garanti. Nous établissons précisément la relation entre ce biais d'estimation sur la variance et l'erreur de Type I résultante des analyses dans le contexte gaussien univarié. Si l'on considère $n$ réalisations d'une variable aléatoire gaussienne $X \sim \mathcal{N}(\mu, \sigma^2)$, l'application de la fission de données sur ces réalisations de $X$ en utilisant un estimateur $\widehat{\sigma^2}$ de $\sigma^2$ pour une classification en deux classes entraîne une déviation de la statistique du $t$-test de Student

sous $\mathcal{H}_0$ égale à :

$$\frac{\sqrt{n}\sqrt{\frac{2}{\pi}\operatorname{Cor}\left(X^{(1)}, X^{(2)}\right)^2}}{\sqrt{1-\frac{2}{\pi}\operatorname{Cor}\left(X^{(1)}, X^{(2)}\right)^2}}. \tag{3}$$

Ce résultat démontre que seul un biais très faible dans l'estimation de $\widehat{\sigma^2}$ est autorisé pour garantir un contrôle de l'erreur de Type I. La Figure 1 **A**, représentant l'évolution de l'erreur de Type I en fonction du biais relatif sur l'estimation de $\sigma^2$, illustre la concordance de ce résultat théorique sur une étude de simulations et souligne ainsi la nécessité d'une méthode d'estimation précise et très efficace de la variance locale.

## 3.2 Performances de l'estimation non-paramétrique de la variance et de son application dans la fission de données

Nous avons réalisé une étude de simulations afin d'évaluer deux aspects cruciaux portant sur notre estimateur de la variance : *i*) ses performances directes en termes de qualité de l'estimation et *ii*) son application dans la fission de données pour des problèmes d'inférence post-classification. Pour ce faire, nous avons généré $n = 100$ réalisations d'un modèle de mélange de gaussiennes univarié à deux composantes ayant des variances égales. Nous avons exploré différentes valeurs de la différence de moyenne $\delta$ entre les deux composantes, allant de $\delta = 0$ (mélange à une seule classe) à $\delta = 100$, ainsi que diverses valeurs de la variance intra-composantes partagée $\sigma^2$. La fission de données a été appliquée à chaque observation en utilisant son estimation non-paramétrique de la variance locale associée, comme décrit en section 2.3. Par la suite, $X^{(1)}$ a été utilisé pour une classification en 3 classes via l'algorithme des $k$-means, induisant ainsi une composante faussement séparée en deux classes. Ensuite, une différence de moyenne entre ces deux classes artificielles a été testée à l'aide du $t$-test de Student sur $X^{(2)}$.

La Figure 1 **B** représente l'évolution du biais relatif sur l'estimation de $\sigma^2$ en fonction de la séparabilité entre les deux composantes du mélange, décrite par le ratio $\delta \,/\, \sigma$. Il est clair d'après cette figure qu'une bonne séparabilité entre les deux classes (une grande valeur du ratio $\delta \,/\, \sigma$) est nécessaire pour que le biais d'estimation soit suffisament faible. La Figure 1 **C** illustre les performances de notre approche en termes de contrôle de l'erreur de Type I avec la fission de données appliquée à l'inférence post-classification. Comme attendu, l'absence de biais dans l'estimation de la variance est cruciale pour garantir un contrôle effectif de l'erreur de Type I. Pour notre méthode d'estimation, cela se traduit donc par la nécessité d'une très bonne séparabilité entre les classes.

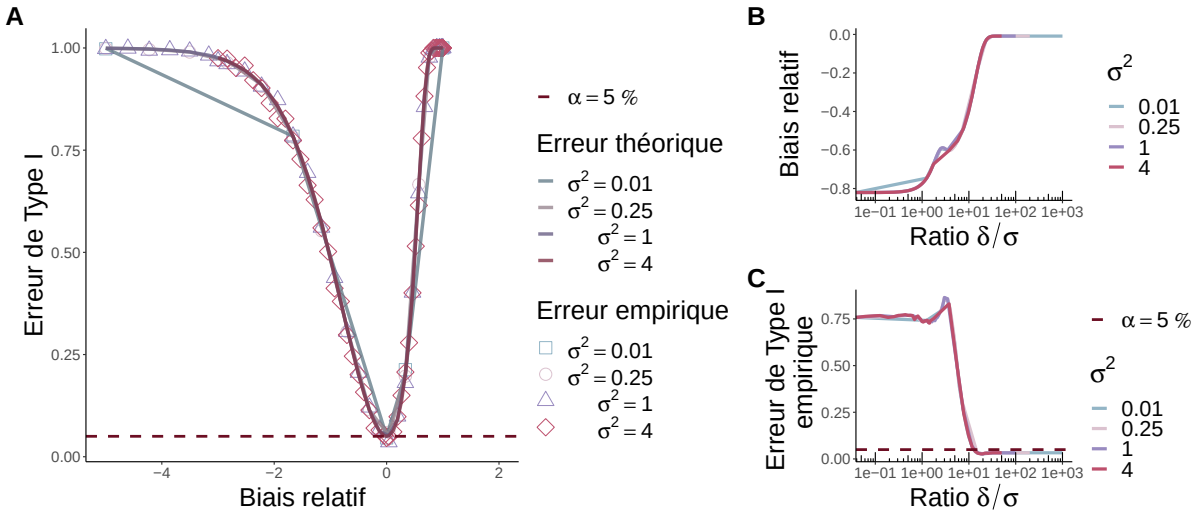



FIGURE 1 : Importance de l'estimation de la variance pour la fission de donnée. **A.** Comparaison de l'erreur de Type I théorique décrite en (3) et de l'erreur de Type I empirique associées à la fission de données pour de l'inférence post-classification en fonction du biais relatif sur l'estimation de $\sigma^2$. **B.** Évolution du biais relatif sur l'estimation de $\sigma^2$ en fonction de la séparabilité entre les classes donnée par le ratio $\delta \,/\, \sigma$. **C.** Évolution de l'erreur du Type I associée à la fission de données en fonction de la séparabilité entre les classes donnée par le ratio $\delta \,/\, \sigma$.

# 4 Discussion

Bien que la fission de données et sa généralisation, le *data thinning*, aient été initialement proposées comme des solutions aux problèmes d'inférence post-classification, ces méthodes se heurtent à des défis majeurs, liés aux hypothèses paramétriques qu'elles exigent. En particulier, l'absence de mélange dans les données et la connaissance préalable des hyperparamètres. Ces conditions restrictives rendent leur application extrêmement difficile en pratique, car la structure sous-jacente des données, essentielle à leur bon fonctionnement, demeure inconnue dans le contexte de la classification non-supervisée.

Nous soulignons ces limites en nous concentrant sur le cas gaussien, où l'estimation de la variance (indispensable à l'utilisation pratique de ces méthodes) devient particulièrement délicate en raison de sa dépendance aux composantes inconnues et à estimer. Nos analyses théoriques ont démontré l'impact significatif d'une mauvaise estimation de la variance sur l'inflation de l'erreur de Type I lors de l'inférence post-classification, mettant en évidence les risques inhérents à l'application aveugle de la fission de données et du *data thining*. Dans le but de surmonter ces limites, nous avons proposé une nouvelle approche reposant sur une estimation non-paramétrique de la variance locale, éliminant ainsi le besoin de connaissance préalable des composantes. Cependant, cette méthode alternative nécessite notamment une très bonne séparabilité entre les vraies composantes afin de garantir un contrôle effectif de l'erreur de Type I. À noter que, bien que nos résultats et analyses soient spécifiques au cas

gaussien, les limites discutées demeurent vraies pour d'autres distributions ayant un processus de fission ou de *data thinning* reposant sur la connaissance d'hyperparamètres. Ainsi, la nécessité de comprendre et de traiter ces limitations devient cruciale lors de l'application de telles méthodes à des données réelles.

# Références


Chen, Y. T. and Gao, L. L. (2023). Testing for a difference in means of a single feature after clustering. *arXiv preprint arXiv:2311.16375*.

Gao, L. L., Bien, J., and Witten, D. (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*.

Hivert, B., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2024). Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, 107916.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540.

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2023). Data fission: splitting a single data point. *Journal of the American Statistical Association*.

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2023). Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*.

Zhang, J. M., Kamath, G. M., and David, N. T. (2019). Valid post-clustering differential analysis for single-cell RNA-Seq. *Cell systems*, 9(4):383–392.

# Mélange de chaînes de Markov d'ordre variable pour l'analyse de séquences

Fabrice Rossi

*CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University,*
*Fabrice.Rossi@dauphine.psl.eu*

**Résumé.**

Nous proposons dans cette communication un modèle de mélange de chaînes de Markov d'ordre variable. Ces chaînes de Markov parcimonieuses permettent l'estimation de dépendances longues dans des séries temporelles à valeurs discrètes, sans pour autant nécessiter des observations longues (en comparaison). Elles sont donc particulièrement adaptées pour modéliser de trajectoires de vie et d'autres processus historiques décrits par un nombre relativement restreint de pas de temps, comme cela est fréquent en sciences humaines. Dans ces domaines, une tâche cruciale est l'identification de groupes de trajectoires présentant des comportements similaires. Nous proposons de réaliser cette tâche au moyen d'un modèle de mélange de chaînes d'ordre variable. L'estimation de ces modèles étant par nature non paramétrique, nous utilisons une vraisemblance pénalisée dont la version complète conduit de façon directe à un algorithme EM. Le nombre de composantes du mélange peut être choisi par l'utilisation de la même vraisemblance pénalisée.

**Mots-clés.** Séries temporelles discrètes, chaînes de Markov d'ordre variable, modèle de mélange, analyse de séquence

**Abstract.** In this talk, we propose a mixture model based variable length Markov chain. These sparse Markov chains allow for the estimation of long dependencies in discrete-valued time series without requiring long observations (in comparison). They are thus particularly suitable for modeling life trajectories and other historical processes described by a relatively small number of time steps, as is common in the social sciences. In these fields, a crucial task is the identification of groups of trajectories exhibiting similar behaviors. We propose to accomplish this task using a variable length Markov chain mixture model. Since the estimation of these models is inherently non-parametric, we employ a penalized likelihood, the complete version of which leads directly to an EM algorithm. The number of mixture components can be chosen using the same penalized likelihood.

**Keywords.** Discrete-valued time series, variable length markov chain, mixture model, sequence analysis

## 1 Introduction

Nous nous intéressons dans cette communication à l'analyse exploratoire de séries temporelles à valeurs discrètes. Cette tâche est, par exemple, particulièrement importante dans

l'analyse de séquences (Liao et al., 2022), une étape fondamentale des recherches sur les trajectoires de vie en sciences sociales. Dans ce contexte, un des objectifs de l'analyse est d'identifier des groupes de trajectoires (donc de séries temporelles à valeurs discrètes) « semblables », qui exhibent le même comportement général.

Historiquement, la première méthode proposée a été celle de l'appariement optimal (*optimal matching*), introduite par Abbott and Forrest (1986), qui s'appuie sur les distances d'édition (insertion, suppression et substitution d'observations pour passer d'une séquence à une autre). Plus généralement, la pratique en sciences sociales est dominée par le calcul de dissimilarités entre séries temporelles (Studer and Ritschard, 2015), mesures qui sont ensuite traitées par une classification hiérarchique ascendante ou par une variante adaptée des *k-means* (Kaufman and Rousseeuw, 1987).

L'utilisation de modèles Markoviens s'est développée plus récemment, notamment au moyen de modèles à ordre variable, les *Variable Length Markov Chain*, VLMC (Bühlmann and Wyner, 1999; Gabadinho and Ritschard, 2016) et des modèles latents, *Hidden Markov Model*, HMM (Bolano and Berchtold, 2016). Le développement de ces approches a naturellement conduit à formuler le problème de classification de trajectoires comme l'estimation d'un mélange de modèles markoviens, notamment dans (Helske and Helske, 2019; Helske et al., 2023).

À notre connaissance, les modèles à ordre variable n'ont jamais été utilisés dans le contexte de la classification. Nous nous proposons donc dans cet article d'étudier les mélanges de VLMC. Nous dérivons en particulier une stratégie d'estimation basée sur une vraisemblance pénalisée optimisée par un algorithme EM.

## 2 Rappel sur les chaînes de Markov d'ordre variable

Les VLMC ont été proposés par Rissanen (1983) et développés notamment par Bühlmann and Wyner (1999). On peut les voir comme des chaînes de Markov d'ordre supérieur parcimonieuses.

Soit $S$ un ensemble fini et $S^\infty$ l'ensemble des séquences de longueurs arbitraires sur $S$. Une série temporelle $(X_i)_{i\in\mathbb{Z}}$ indexée par les entiers relatifs et à valeurs dans $S$ est un VLMC s'il existe une fonction $l$ de $S^\infty$ dans $\{0,\dots,l_{\max}\}$ telle que pour tout $t$ et toute séquence $x_{-\infty}^t$

$$\mathbb{P}(X_t = x_t \mid X_{-\infty}^{t-1} = x_{-\infty}^{t-1}) = \mathbb{P}\left(X_t = x_t \mid X_{t-l(x_{-\infty}^t)}^{t-1} = x_{t-l(x_{-\infty}^t)}^{t-1}\right). \tag{1}$$

Dans cette définition on a utilisé la notation suivante : pour toute séquence $(t_i)_i$ indexée sur $\mathbb{N}$ ou $\mathbb{Z}$, $t_i^j$ désigne la séquence $(t_i, t_{i+1}, \dots, t_j)$, avec $i < j$. En particulier $t_{-\infty}^i$ désigne la séquence infinie à gauche.

On voit qu'un VLMC est donc une chaîne de Markov d'ordre $l_{\max}$ parcimonieuse. En effet, sa mémoire, et donc le besoin de définir une probabilité conditionnelle, dépend du *contexte*. La fonction contexte correspondante, $c$, de $S^\infty$ dans lui-même, associe à un passé quelconque $x_{-\infty}^t$ le passé « important », le contexte, $x_{t-l(x_{-\infty}^t)}^{t-1}$. Au lieu de spécifier $S^{l_{\max}}$ lois

conditionnelles, on se contente d'autant de lois qu'il existe de contextes différents (l'image de $S^\infty$ par $c$). On découple ainsi $l_{\max}$ du nombre de paramètres du modèle et on évite l'explosion de ce dernier avec la longueur de la mémoire. Il ainsi possible d'observer des dépendances temporelles longues conjointement à des dépendances courtes.

# 3 Mélange de VLMC

## 3.1 Mélange de chaînes de Markov

Un modèle VLMC peut être vu comme une chaîne de Markov d'ordre $l_{\max}$ dans laquelle on force les lois conditionnelles de même contexte à être identique. De ce fait, un mélange de VLMC est un mélange de chaînes de Markov (cf par exemple van de Pol and Langeheine (1990)), au moins d'un point de vue superficiel.

Concrètement, on suppose qu'on observe des séries temporelles à valeurs dans l'ensemble fini $S$, engendrées par un mélange de $K$ VLMC, $\mathcal{M}^1, \ldots, \mathcal{M}^K$. Le mélange est caractérisé par des probabilités *a priori* $\pi_1, \ldots, \pi_K$. À chaque série temporelle $X^i = (X^i_{1 \leq t \leq m_i})$, on associe une variable latente $Z^i$ distribuée selon $\pi$ (i.e. $\mathbb{P}(Z^i = k \mid \pi) = \pi_k$). Les $Z^i$ sont indépendantes, de même que les $X^i$. $X^i$ est engendrée par $\mathcal{M}^k$ si $Z^i = k$. On suppose les $m_i$ déterministes.

Étant données $N$ séries temporelles, on cherche à estimer $\pi$ et les $K$ VLMC, c'est-à-dire pour ces derniers les fonctions de contexte $(c^k)_{1 \leq k \leq K}$ et les probabilités conditionnelles associées.

## 3.2 Estimation d'un VLMC

L'estimation d'un VLMC à partir d'une série temporelle se fait classiquement à partir de l'algorithme contexte de Rissanen (1983). Il identifie les sous-séquences qui apparaissent dans la série et dont l'interprétation comme un contexte apporte assez d'information au sens d'un test de rapport de vraisemblance (cf par exemple Bühlmann and Wyner (1999) pour des détails). Les probabilités conditionnelles sont estimées au sens du maximum de vraisemblance, mais ce n'est pas le cas de la fonction de contexte elle-même. Comme l'observe par exemple Mächler and Bühlmann (2004), la complexité du modèle (mesurée par son nombre de contextes) dépend fortement du niveau retenu pour le test.

Dans le cadre un mélange de VLMC, il est donc préférable de recourir à une estimation par vraisemblance pénalisée, comme proposé (pour un seul modèle) dans Csiszar and Talata (2006); Garivier (2006); Garivier and Leonardi (2011). Si on note $|\mathcal{M}^k|$ le nombre de contextes du VLMC $\mathcal{M}^k$, une pénalité naturelle est $|\mathcal{M}^k| f(M)$ où $M$ est le nombre total d'observations (ici $M = \sum_{i=1}^N m_i$) et $f$ une fonction positive choisie comme dans Garivier and Leonardi (2011) : $f(m) \to \infty$ et $\frac{f(m)}{m} \to 0$ quand $m \to \infty$. Un exemple classique est la pénalité du BIC, avec $f(m) = \frac{|S|-1}{2} \log m$ (où $|S|$ est le cardinal de l'espace d'états, i.e. le nombre de valeurs discrètes possibles pour les séries observées).

## 3.3 Algorithme EM

Notons $p(\mathcal{M}^k \mid x^i)$ la vraisemblance du VLMC $\mathcal{M}^k$ pour la série $x^i$. La log-vraisemblance complète pénalisée du mélange de $K$ VLMC s'écrit alors

$$L_p(\pi,\ \mathcal{M}^1, \ldots, \mathcal{M}^K \mid \mathbf{z}, \mathbf{x}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_k^i \left(\log \pi_k + \log p(\mathcal{M}^k \mid x^i)\right) - f(M) \sum_{k=1}^{K} |\mathcal{M}^k|,$$

où $\mathbf{z} = (z^1, \ldots, z^N)$, $\mathbf{x} = (x^1, \ldots, x^N)$ et où les $z_k^i$ sont les indicatrices $z_k^i = \mathbb{I}_{z^i=k}$.

Cette formulation conduit de façon directe à un algorithme EM. Dans la phase E de l'algorithme, on calcule

$$\mathbb{E}_{\mathbf{Z}\sim q}\left\{L_p(\pi,\ \mathcal{M}^1, \ldots, \mathcal{M}^K \mid \mathbf{Z}, \mathbf{x})\right\} = \mathbb{E}_{\mathbf{Z}\sim q}\left\{\sum_{i=1}^{N} \sum_{k=1}^{K} z_k^i \left(\log \pi_k + \log p(\mathcal{M}^k \mid x^i)\right)\right\} - f(M) \sum_{k=1}^{K} |\mathcal{M}^k|$$

où $q$ désigne une distribution sur les variables latentes $\mathbf{Z} = (Z^1, \ldots, Z^N)$. Si l'estimation courante des paramètres est $\pi^{(t)}$, $\mathcal{M}^{1(t)}, \ldots, \mathcal{M}^{K(t)}$, la distribution optimale pour les variables latentes est donnée par

$$q(Z^i = k)^{(t)} = \tau_k^{i(t)} = \mathbb{P}(Z^i = k \mid x^i,\ \pi^{(t)},\ \mathcal{M}^{1(t)}, \ldots, \mathcal{M}^{K(t)}),$$

c'est-à-dire

$$\tau_k^{i(t)} = \frac{\pi_k^{(t)} \mathbb{P}(x^i \mid \mathcal{M}^{k(t)})}{\sum_{l=1}^{K} \pi_l^{(t)} \mathbb{P}(x^i \mid \mathcal{M}^{l(t)})}.$$

On a donc

$$\mathbb{E}_{\mathbf{Z}\sim q^{(t)}}\left\{L_p(\pi,\ \mathcal{M}^1, \ldots, \mathcal{M}^K \mid \mathbf{Z}, \mathbf{x})\right\} = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_k^{i(t)} \left(\log \pi_k + \log p(\mathcal{M}^k \mid x^i)\right) - f(M) \sum_{k=1}^{K} |\mathcal{M}^k|$$

Dans la phase $M$ on maximise cette quantité par rapport à $\pi$ et aux VLMC $\mathcal{M}^1, \ldots, \mathcal{M}^K$. Or, il est clair que les différents VLMC peuvent être optimisés séparément en maximisant pour chaque $\mathcal{M}^k$

$$\sum_{i=1}^{N} \tau_k^{i(t)} \log p(\mathcal{M}^k \mid x^i) - f(M)|\mathcal{M}^k|.$$

En pratique, l'implémentation de la phase M se base donc sur l'algorithme proposé dans Garivier (2006) qui peut être lui-même vu comme une variante de l'algorithme contexte. Deux adaptations sont nécessaires : la prise en compte de plusieurs séries temporelles et l'intégration d'une pondération pour chaque série.

Notons pour conclure que le nombre de composantes sur mélange peut être lui-même choisi à partir de la vraisemblance pénalisée.

# Bibliographie


A. Abbott and J. Forrest. Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, 16(3) :471–494, 1986. doi : 10.2307/204500.

D. Bolano and A. Berchtold. General framework and model building in the class of hidden mixture transition distribution models. *Computational Statistics & Data Analysis*, 93 : 131–145, 2016. ISSN 0167-9473. doi : https://doi.org/10.1016/j.csda.2014.09.011. URL https://www.sciencedirect.com/science/article/pii/S0167947314002722.

P. Bühlmann and A. J. Wyner. Variable length markov chains. *The Annals of Statistics*, 27 (2) :480–513, April 1999. doi : 10.1214/aos/1018031204.

I. Csiszar and Z. Talata. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information Theory*, 52(3) :1007–1016, March 2006. ISSN 1557-9654. doi : 10.1109/TIT.2005.864431.

A. Gabadinho and G. Ritschard. Analyzing state sequences with probabilistic suffix trees : The pst r package. *Journal of Statistical Software*, 72(3) :1–39, 2016. doi : 10.18637/jss.v072.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v072i03.

A. Garivier. Consistency of the unlimited bic context tree estimator. *IEEE Transactions on Information Theory*, 52(10) :4630–4635, Oct 2006. ISSN 1557-9654. doi : 10.1109/TIT.2006.881742.

A. Garivier and F. Leonardi. Context tree selection : A unifying view. *Stochastic Processes and their Applications*, 121(11) :2488–2506, 2011. ISSN 0304-4149. doi : https://doi.org/10.1016/j.spa.2011.06.012.

S. Helske and J. Helske. Mixture hidden markov models for sequence data : The seqhmm package in r. *Journal of Statistical Software*, 88(3) :1–32, 2019. doi : 10.18637/jss.v088.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v088i03.

S. Helske, M. Keski-Säntti, J. Kivelä, A. Juutinen, A. Kääriälä, M. Gissler, M. Merikukka, and T. Lallukka. Predicting the stability of early employment with its timing and childhood social and health-related predictors : a mixture markov model approach. *Longitudinal and Life Course Studies*, 14(1) :73 – 104, 2023. doi : 10.1332/175795921X16609201864155.

L. Kaufman and P. J. Rousseeuw. Clustering by means of medoids. In Y. Dodge, editor, *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 405–416. North-Holland, 1987.

T. F. Liao, D. Bolano, C. Brzinsky-Fay, B. Cornwell, A. E. Fasang, S. Helske, R. Piccarreta, M. Raab, G. Ritschard, E. Struffolino, and M. Studer. Sequence analysis : Its past, present, and future. *Social Science Research*, 107 :102772, 2022. doi : 10.1016/j.ssresearch.2022.102772.

M. Mächler and P. Bühlmann. Variable length markov chains : Methodology, computing, and software. *Journal of Computational and Graphical Statistics*, 13(2) :435–455, 2004. doi : 10.1198/1061860043524.

J. Rissanen. A universal data compression system. *IEEE Transactions on Information Theory*, 29(5) :656–664, Sep. 1983. ISSN 1557-9654. doi : 10.1109/TIT.1983.1056741.

M. Studer and G. Ritschard. What Matters in Differences Between Life Trajectories : A Comparative Review of Sequence Dissimilarity Measures. *Journal of the Royal Statistical Society Series A : Statistics in Society*, 179(2) :481–511, 07 2015. ISSN 0964-1998. doi : 10.1111/rssa.12125.

F. van de Pol and R. Langeheine. Mixed markov latent class models. *Sociological Methodology*, 20 :213–247, 1990.

# Extrêmes et risques

# Intervalles de confiance pour la prime de réassurance en présence de risques extrêmes

Abdelkader Ameraoui [1] & Kamal Boukhetala [2] & Jean-François Dupuy [3]

[1] *Ecole Nationale Supérieure de Mathématiques, Alger, Algérie, a.ameraoui@gmail.com*
[2] *Université des Sciences et de la Technologie Houari Boumédiène, Alger, Algérie, kboukhetala@usthb.dz*
[3] *Université de Rennes, IRMAR-INSA, France, jean-francois.dupuy@insa-rennes.fr*

**Résumé.** Nous nous intéressons à la construction d'intervalles de confiance pour la prime de réassurance en présence de risques extrêmes. Une première méthode de construction d'intervalles de confiance, simple à mettre en œuvre, repose sur la normalité asymptotique d'un estimateur de la prime, proposé par Necir et al. (2007). Nos simulations sur des échantillons de taille finie montrent néanmoins que la probabilité de couverture de ces intervalles peut être très inférieure au niveau nominal. Dans cet exposé, nous proposons donc deux autres méthodes de construction d'intervalles de confiance pour la prime de réassurance en présence de risques exrêmes. La première est basée sur un rapport de vraisemblance, la seconde sur une méthode de pondération des observations dans un calcul de vraisemblance sous contraintes. Nous en déduisons deux variables asymptotiquement pivotales, qui nous permettent de construire des intervalles de confiance asymptotiques. Ces deux méthodes, ainsi que la méthode basée sur l'estimateur de Necir et al. (2007), sont évaluées par simulations puis illustrées sur un jeu de données contenant les montants de sinistres incendie. La méthode basée sur la pondération des observations dans un calcul de vraisemblance sous contraintes apparaît comme la plus performante, en termes de probabilité de couverture et de longueur des intervalles.

**Mots-clés.** Estimateur de Hill, rapport de vraisemblance, méthode data tilting

**Abstract.** We consider the construction of confidence intervals for the reinsurance premium of extreme risks. A straightforward method is based on the asymptotic normality of an estimator of the premium proposed by Necir et al. (2007). However, our simulations suggest that the coverage accuracy of the resulting intervals can be quite far from the nominal confidence level. Therefore in this talk, we propose two alternative methods. One is based on a likelihood ratio, the other on a data tilting method (data tilting amounts to a constrained minimisation problem of some distance function). These methods and the method based on the asymptotic normality of Necir et al. (2007) estimator are evaluated via simulations and illustrated on a real data set of fire loss. The data tilting method appears to be the most efficient, in terms of coverage probabilities and intervals length.

**Keywords.** Hill estimator, likelihood ratio method, data tilting method

# 1 Objectives

Extreme events, which arise in a wide variety of domains (e.g., environment, industry, finance) can cause considerable loss in insurers portfolio. It is thus crucial for (re)-insurance companies to determine adequate premiums for extreme risks.
One common premium calculation principle, due to Wang (1996), is based on a distortion function, that is, an increasing and concave function $g : [0, 1] \to [0, 1]$ such that $g(0) = 0$ and $g(1) = 1$. If $X$ is a random risk with distribution function $F$, the distortion risk measure based on $g$ is defined as

$$\Pi(g) = \int_0^\infty g(1 - F(x))\, dx.$$

Several distorsion functions have been proposed. Letting $g(x) = x^{1/\rho}$, for $\rho \geq 1$, yields the popular Proportional Hazard (PH) transform. The PH premium of $X$ is given by:

$$\Pi_\rho = \int_0^\infty (1 - F(x))^{\frac{1}{\rho}}\, dx,$$

which can be seen as a distorted expectation of $X$ (the parameter $\rho$ controls the amount of risk loading in the premium, and is called the risk aversion index).
Now, in reinsurance, a standard practice is that the reinsurer compensates the cedant's loss only above a certain retention amount $R > 0$. In this case, the reinsurer will not pay the insurer if $X$ is less than or equal to $R$ and will pay $(X - R)$ if $X$ exceeds $R$. The amount paid by the reinsurer is thus $(X - R)_+$, where $x_+ = \max(0, x)$, and the corresponding PH premium for the layer $[R, \infty)$ is defined as the distorted expectation of $(X - R)_+$:

$$\Pi_{\rho,R} = \int_R^\infty (1 - F(x))^{\frac{1}{\rho}}\, dx.$$

Various estimates of $\Pi_{\rho,R}$ have been proposed for heavy-tailed insured risk, see for example Necir and Boukhetala (2004), Necir et al. (2007), Benkhelifa (2014), Ahmedou et al. (2023).

Our objective, here, is to construct confidence intervals for $\Pi_{\rho,R}$. In Necir et al. (2007), an asymptotically normal estimate of $\Pi_{\rho,R}$ is proposed. From this, it is straigthforward to construct a confidence interval for $\Pi_{\rho,R}$. However, our simulation study (Ameraoui et al., 2023) suggests that the coverage probabilities of the resulting intervals can be quite far from the nominal confidence level. Thus, in this work, we consider two alternative methods, namely: *i*) a likelihood ratio method, and *ii*) a data tilting method. Both have already proved useful to construct confidence intervals for the tail index and high quantiles of a heavy-tailed distribution, see for example Lu and Peng (2002), Peng and Qi (2006), Tursunalieva and Silvapulle (2016). We adapt them to the setting of interval estimation of the PH premium under high-excess loss layer.

# 2 Results

We propose two asymptotically pivotal functions for $\Pi_{\rho,R}$, based on the likelihood ratio and data tilting methods, and we prove their convergence in distribution to a $\chi^2_1$ distribution.

From these results, we can construct confidence intervals for $\Pi_{\rho,R}$. Their performance are assessed in a simulation study. Evaluation criteria include coverage probabilities and interval length.
Our results (see Ameraoui et al., 2023) suggest that the data tilting method provides the best results. The confidence intervals based on this method have a coverage accuracy close to the nominal level, and their length are generally smaller. They are also less sensitive to the choice of the sample fraction used to calculate the various quantities involved in the intervals (such as the Hill estimate of the tail index of $F$).

# Bibliographie


Ahmedou, S., Deme, E.H. and Fofana, S. (2023), An improved estimator of distortion risk premiums under dependence insured risks with heavy-tailed marginals, *Far East Journal of Theoretical Statistics*, 67(3), pp. 243-277.

Ameraoui, A., Boukhetala, K. and Dupuy, J.-F. (2023), Confidence intervals for the proportional hazard reinsurance premium for heavy-tailed claims, submitted.

Benkhelifa, L. (2014), Kernel-type estimator of the reinsurance premium for heavy-tailed loss distributions, *Insurance: Mathematics and Economics* 59, pp. 65-70.

Lu, J.C. and Peng, L. (2002), Likelihood based confidence intervals for the tail index, *Extremes* 5, pp. 337-352.

Necir, A. and Boukhetala, K. (2004), Estimating the risk adjusted premium of the largest reinsurance covers, *In: Antoch, Jaromir (Ed.), Proceeding of Computational Statistics. Physica-Verlag, Springer*, pp. 1577-1584.

Necir, A., Meraghni, D. and Meddi, F. (2007), Statistical estimate of the proportional hazard premium of loss, *Scandinavian Actuarial Journal*, 3, pp. 147-161.

Peng, L. and Qi, Y. (2006), Confidence regions for high quantiles of a heavy tailed distribution, *The Annals of Statistics*, 34(4), pp. 1964-1986.

Tursunalieva, A. and Silvapulle, P. (2016), Nonparametric estimation of operational value-at-risk (OpVaR), *Insurance: Mathematics and Economics*, 69, pp. 194-201.

Wang, S. (1996), Premium calculation by transforming the layer premium density, *Astin Bulletin*, 26(1), pp. 71-92.

# Reduction de dimension pour l'estimation de l'indice des valeurs extrêmes conditionnel

Laurent Gardes [1] & Alex Podgorny [2]

[1] *Université de Strasbourg, CNRS, IRMA UMR 7501, F-67000 Strasbourg, France. E-mail :* gardes@unistra.fr
[2] *Université de Strasbourg, CNRS, IRMA UMR 7501, F-67000 Strasbourg, France. E-mail :* apodgorny@unistra.fr

**Résumé.** Nous nous intéressons à la relation liant les grandes valeurs d'une variable aléatoire réelle $Y$ à une covariable $X$ prenant ses valeurs dans un sous ensemble $\mathcal{X}$ de $\mathbb{R}^p$. Pour ce faire, nous supposons que la loi conditionnelle de $Y$ sachant $X = x$ est une loi à queue lourde d'indice des valeurs extrêmes $\gamma(x)$. L'estimation de cet indice est une étape essentielle pour l'inférence de la loi conditionnelle mais cette tâche est d'autant plus délicate que la dimension $p$ augmente. L'objectif de ce travail est de proposer une méthode de réduction de dimension dans le but d'obtenir un estimateur plus efficace de l'indice des valeurs extrêmes. Plus précisément, nous supposerons qu'il existe un sous-espace $\mathcal{S}_0$ de dimension $q < p$ de base $B_0 \in \mathbb{R}^{p\times q}$ et une fonction positive $g(\cdot)$ tels que pour tout $x \in \mathcal{X}$, $\gamma(x) = g(B_0^\top x)$. Nous proposons une méthode d'estimation de ce sous-espace et en établissons la consistance. Nous illustrons les avantages de cette procédure de réduction de dimension pour l'estimation de l'indice de valeurs extrêmes conditionnel à l'aide de simulations.
**Mots-clés.** Indice des valeurs extrêmes, réduction de dimension, lois à queue lourde.

**Abstract.** We are interested in the relationship between the large values of a real random variable $Y$ and its associated covariate $X$ that takes its values in a subset $\mathcal{X}$ of $\mathbb{R}^p$. To do this, we assume that the conditional distribution of $Y$ given $X = x$ has a heavy-tailed distribution with tail index $\gamma(x) > 0$. Estimating this index is a crucial step for the inference of the conditional distribution, but this task becomes more challenging as the dimension $p$ increases. The objective of this work is to propose a dimension reduction method to obtain a more efficient estimator of the extreme value index. Specifically, we assume the existence of a subspace $\mathcal{S}_0$ of dimension $q < p$ with basis $B_0 \in \mathbb{R}^{p\times q}$ and a positive function $g(\cdot)$ such that for all $x \in \mathcal{X}$, $\gamma(x) = g(B_0^\top x)$. We propose a method to estimate this subspace and establish its consistency. We illustrate the advantages of this dimension reduction procedure for estimating the extreme value index through simulations.
**Keywords.** Tail-index, dimension reduction, heavy-tailed distrbutions.

# 1 Introduction

On considère un couple aléatoire $(X, Y)$ où $Y$ est une variable aléatoire positive et $X$ une variable aléatoire de support $\mathcal{X} \subset \mathbb{R}^p$. Nous nous plaçons dans le cas où pour tout $x \in \mathcal{X}$, la loi conditionnelle de $Y$ sachant $X = x$ est à queue lourde d'indice des valeurs extrêmes conditionnel $\gamma(x) > 0$. Cet indice contrôle le comportement de la queue de la loi conditionnelle. Son estimation est essentielle dans de nombreux domaines comme en finance (voir, par exemple, Rockafellar et Uryasev [8]) ou en assurance (voir par exemple Brazauskas *et al.* [3] et Read et Vogel [7]. Plusieurs estimateurs de l'indice des valeurs extrêmes conditionnel sont disponibles dans la littérature citons par exemple Daouia, et al. [4], Gardes et Stupfler [5] ou encore Goegebeur et al. [6]. Cependant, il est bien connu que, pour un niveau de précision donné de l'un de ces estimateurs, le nombre $n$ d'observations croît de manière exponentielle avec la dimension $p$. Ce phénomène est souvent appelé le fléau de la dimension (voir Bellman [2]).
Le point de départ de ce travail est de supposer l'existence d'un sous espace $\mathcal{S}$ de base $B \in \mathbb{R}^{p\times q}$ tel que, pour tout $x \in \mathcal{X}$, $\gamma(x) = g(B^\top x)$, où $g(\cdot)$ est une fonction positive inconnue. Ce sous espace est appelé sous-espace TIDR (pour Tail-index dimension reduction en anglais). Par conséquent, si la matrice $B$ est connue (ou peut au moins être estimée), l'estimation de $\gamma(X)$ peut être effectuée en remplaçant $X$ par la covariable $B^\top X$ de dimension réduite $q \leq p$.

La principale contribution de ce travail est l'introduction d'un nouveau sous-espace DR, appelé , qui est utile lorsque nous cherchons à estimer l'indice des valeurs extrêmes conditionnel. Nous proposons ensuite une procédure d'estimation de ce sous-espace et nous démontrons sa consistance.

## 1.1 Définition du sous-espace TIDR

Dans la suite, nous considérons le modèle suivant pour la loi du couple $(X, Y)$.

(M) Le support $\mathcal{X}$ de $X$ est supposé compact avec un intérieur non vide et l'extrémité gauche de la distribution de $Y$ est strictement positive. De plus, pour tout $x \in \mathcal{X}$, la distribution conditionnelle de $Y$ sachant $X = x$ est à queue lourde avec un indice de queue $\gamma(x) > 0$.

Rappelons que la loi de $Y$ sachant $X = x$ est à queue lourde si

$$S(y, x) = y^{-1/\gamma(x)}\mathcal{L}(y, x),$$

où $\mathcal{L}(\cdot, x)$ est une fonction à variations lentes c'est-à-dire telle que pour tout $t > 0$ et $x \in \mathcal{X}$,

$$\lim_{y\to\infty} \frac{\mathcal{L}(ty, x)}{\mathcal{L}(y, x)} = 1.$$

Un premier résultat important en vue de la définition du sous-espace TIDR est le suivant. Pour une matrice donnée $B \in \mathbb{R}^{p\times q}$ avec $1 \leq q < p$, sous le modèle (M) et sous certaines hypothèses de régularité que nous ne détaillons

pas ici, la loi conditionnelle de $Y$ sachant $B^\top X = B^\top x$ est également à queue lourde d'indice des valeurs extrêmes

$$\xi_B(B^\top x) := \max_{z: B^\top z = B^\top x} \gamma(z).$$

Une explication intuitive de ce résultat est la suivante. Notons

$$S_B(y, B^\top x) := \mathbb{P}(Y > y \mid B^\top X = B^\top x),$$

la fonction de survie de $Y$ sachant $B^\top X = B^\top x$. La fonction $S_B(y, \cdot)$ peut être vue comme un mélange de la fonction $S(y, \cdot)$ en ce sens que $S_B(y, B^\top X) = \mathbb{E}\left[S(y, X) | B^\top X\right]$. Par conséquent, pour tout $x \in \mathcal{X}$, il est naturel de penser que la vitesse de décroissance de $S_B(\cdot, x)$ est donnée par le plus grand indice des valeurs extrêmes impliqué dans le mélange.

Un sous-espace TIDR est alors défini de la façon suivante.

**Définition 1.** *Un sous-espace vectoriel $\mathcal{S}$ de dimension $q \in \{1, \ldots, p\}$ avec $B \in \mathbb{R}^{p \times q}$ est un sous-espace TIDR si $\xi_B(B^\top X) = \gamma(X)$ presque sûrement.*

Lorsque $B$ est la matrice identité de dimension $p$, il est facile de voir que $\xi_B(B^\top x) = \gamma(x)$ pour tout $x \in \mathcal{X}$ et donc que $\mathcal{S} = \mathbb{R}^p$ est toujours un sous-espace TIDR. Bien entendu, nous cherchons à trouver le plus petit sous-espace.

**Définition 2.** *Un sous-espace linéaire $\mathcal{S}_0$ est le sous-espace central tail index (CTI) si $\mathcal{S}_0$ est un sous-espace TIDR tel que $\mathcal{S}_0 \subset \mathcal{S}$ pour tous les sous-espaces TIDR $\mathcal{S}$.*

La base $B \in \mathbb{R}^{p \times q}$ telle que $\mathcal{S} = \text{span}(B)$ n'est bien sûr pas unique. Il est plus commode, en particulier pour les besoins de l'estimation, de travailler avec la base canonique de $\mathcal{S}$. L'ensemble des bases canoniques des sous espaces de dimension $q$ est noté $\mathcal{B}_q$.

Dans ce qui suit, nous désignons par $B_0 \in \mathcal{B}_q$ la base canonique du sous-espace CTI $\mathcal{S}_0$. La distribution conditionnelle de $Y$ sachant $B_0^\top X$ (qui n'est pas nécessairement égale à celle de $Y$ étant donné $X$) est à queue lourde d'indice $\xi_{B_0}(B_0^\top X) = \gamma(X)$. Par conséquent, en supposant que $B_0$ est connu, l'indice des valeurs extrêmes conditionnel peut être estimé en utilisant un échantillon du couple aléatoire $(B_0^\top X, Y) \in \mathbb{R}^q \times \mathbb{R}$ au lieu de $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$. Lorsque $q \ll p$, cela nous permet de construire un estimateur plus efficace de $\gamma(\cdot)$.

## 1.2 Estimation du CTI

Dans cette partie, nous supposons qu'il existe un sous-espace CTI de dimension connue $q$. Pour estimer la base canonique $B_0 \in \mathcal{B}_q$ du sous-espace CTI, nous commençons par remarquer que pour tout $x \in \mathcal{X}$ et $B \in \mathcal{B}_q$, on a

$$\gamma(x) = \xi_{B_0}(B_0^\top x) \leq \xi_B(B^\top x),$$

puisque $\gamma(x)$ appartient au support de la distribution conditionnelle de $\gamma(X)$ donnée par $B^\top X = B^\top x$. Ainsi, pour tout $B \in \mathcal{B}_q$ avec $B \neq B_0$, $\mathbb{P}\left(\xi_{B_0}(B_0^\top X) < \xi_B(B^\top X)\right) > 0$, ce qui conduit à

$$\arg\min_{B \in \mathcal{B}_q} \mathbb{E}\left[\xi_B(B^\top X)\right] = \{B_0\}.$$

En particulier, soit $\mathcal{X}_0$ un sous-ensemble compact à l'intérieur de $\mathcal{X}$ satisfaisant

(C) pour tout $B \in \mathcal{B}_q$, $B \neq B_0$,

$$\mathbb{P}\left[\left\{\xi_{B_0}(B_0^\top X) < \xi_B(B^\top X)\right\} \cap \{X \in \mathcal{X}_0\}\right] > 0.$$

Nous avons alors

$$\arg\min_{B \in \mathcal{B}_q} \mathbb{E}\left[\xi_B(B^\top X)\mathbb{I}_{\mathcal{X}_0}(X)\right] =: \arg\min_{B \in \mathcal{B}_q} \Psi(B, \mathcal{X}_0) = \{B_0\}.$$

Remarquez que la condition (C) implique que le sous-espace CTI $\mathcal{S}_0$ existe. L'indicatrice $\mathbb{I}_{\mathcal{X}_0}(X)$ est introduite ici pour garantir que la densité de $X$ sur $\mathcal{X}_0$ soit suffisamment éloignée de zéro.
Étant donné un échantillon $(X_1, Y_1), \cdots, (X_n, Y_n)$ de copies indépendantes du couple aléatoire $(X, Y)$, la première étape de la procédure d'estimation de $B_0$ est l'estimation de la fonction $B \mapsto \Psi(B, \mathcal{X}_0)$. La version empirique de $\Psi(B, \mathcal{X}_0)$ est donnée par

$$\widetilde{\Psi}_n(B, \mathcal{X}_0) := \frac{1}{n}\sum_{i=1}^{n} \xi_B(B^\top X_i)\mathbb{I}_{\mathcal{X}_0}(X_i).$$

Nous devons ensuite trouver un estimateur approprié de $\xi_B(B^\top x)$. Pour motiver la définition de notre estimateur, commençons par la situation irréaliste où nous disposons d'un échantillon $\{\check{Z}_i(B, x),\ i = 1, \cdots, m\}$ de $m \in \mathbb{N} \setminus \{0\}$ variables aléatoires indépendantes ayant pour loi commune la loi conditionnelle de $Y$ sachant $B^\top X = B^\top x$. Les statistiques d'ordre associées sont notées $\check{Z}_{(1)}(B, x) < \cdots < \check{Z}_{(n)}(B, x)$. L'estimateur le plus connu de l'indice des valeurs extrêmes est l'estimateur de Hill, qui est donné par la formule suivante

$$\check{\xi}_B^{(H)}(B^\top x) := \frac{1}{\lfloor \alpha m \rfloor}\sum_{i=1}^{\lfloor \alpha m \rfloor} \ln \frac{\check{Z}_{(m-i+1)}(B, x)}{\check{Z}_{(m-\lfloor \alpha m \rfloor)}(B, x)},$$

pour tout $\alpha \in ]1/m, 1[$. Bien entendu, il ne s'agit pas d'un estimateur puisque les variables aléatoires $\{\check{Z}_i(B, x),\ ; i = 1, \cdots, m\}$ ne sont pas observées. Nous proposons donc de les remplacer dans l'expression de $\check{\xi}_B^{(H)}(B^\top x)$ par un ensemble de variables aléatoires observées. Plus précisément, introduisons l'ensemble

$$\mathcal{T}(B, x, h) := \{z \in \mathcal{X} \mid \|B^\top z - B^\top x\| \leq h\},$$

où $h = h_n > 0$ et

$$M^\star = M(B, x, h) := \sum_{i=1}^{n} \mathbb{I}_{\mathcal{T}(B,x,h)}(X_i),$$

le nombre aléatoire de covariables dans l'ensemble $\mathcal{T}(B,x,h)$. Nous notons par $\{W_i^\star = W_i(B,x,h),\ i=1,\cdots,M^\star\}$ l'ensemble des covariables qui appartiennent à $\mathcal{T}(B,x,h)$. Les variables $Y_i$ associées sont notées $\{Z_i^\star = Z_i(B,x,h),\ i=1,\cdots,M^\star\}$. Pour alléger les notations, nous mettons une étoile $(\star)$ pour rappeler la dépendance en $B$, $x$ et $h$. Intuitivement, lorsque $h$ est suffisamment proche de zéro, la variable aléatoire $Z_i^\star$ est approximativement distribuée comme $\check{Z}_i(B,x)$. Pour une suite $(\alpha_n) = (\alpha) \in ]0,1[$, cela nous amène à introduire l'estimateur de l'indice des valeurs extrêmes conditionnel défini ci-dessous.

**Définition 3.** *Sous le modèle (M), pour tout $(B,x) \in \mathcal{B}_q \times \mathcal{X}$, l'estimateur local de Hill de l'indice de queue $\xi_B(B^\top x)$ est*

$$\widehat{\xi}_B^{(H)}(B^\top x) = \widehat{\xi}_B^{(H)}(B^\top x, \alpha, h) := \frac{1}{\lfloor \alpha M^\star \rfloor} \sum_{i=1}^{\lfloor \alpha M^\star \rfloor} \ln \frac{Z^\star_{(M^\star - i+1)}}{Z^\star_{(M^\star - \lfloor \alpha M^\star \rfloor)}},$$

*si $\alpha M^\star > 1$ et $\widehat{\xi}_B^{(H)}(B^\top x, \alpha, h) = 0$ sinon.*

Nous proposons ainsi d'estimer $\Psi(B, \mathcal{X}_0)$ pour tout $B \in \mathcal{B}_q$ par l'estimateur plug-in

$$\widehat{\Psi}_n^{(H)}(B, \mathcal{X}_0) := \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_B^{(H)}(B^\top X_i) \mathbb{I}_{\mathcal{X}_0}(X_i). \tag{1}$$

La définition de l'estimateur de $B_0$ est donnée ci-dessous.

**Définition 4.** *Sous le modèle (M), pour un échantillon $(X_1,Y_1),\cdots,(X_n,Y_n)$ de copies indépendantes du couple aléatoire $(X,Y)$, l'estimateur $\widehat{B}_n$ de la base canonique du sous-espace CTI minimise la fonction $B \mapsto \widehat{\Psi}_n(B, \mathcal{X}_0)$.*

## 1.3 Consistance de l'estimateur $\widehat{B}_n$.

Sous des hypothèses supplémentaires techniques que nous ne mentionnons pas ici, nous montrons que

$$\left\| \widehat{B}_n - B_0 \right\| \xrightarrow{\mathbb{P}} 0$$

où $\|\cdot\|$ est une norme quelconque dans $\mathbb{R}^{p\times q}$.
La démonstration de ce résultat passe essentiellement par celle de la consistance uniforme de l'estimateur $\widehat{\Psi}_n^{(H)}(B, \mathcal{X}_0)$ de $\Psi(B, \mathcal{X}_0)$ i.e.,

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}_0} \sup_{B\in\mathcal{B}_q} \left| \widehat{\Psi}_n^{(H)}(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right| = 0.$$

## 1.4 Simulations

On se place dans le cadre suivant. La variable aléatoire $X$ suit une loi uniforme sur l'espace $\mathcal{X} := [0,1]^p$ que l'on munit de la norme uniforme $\|\cdot\|_\infty$. Nous fixons

la dimension de la covariable $X$ à $p = 8$. La variable aléatoire $Y$ est donnée par $Y := Q(U, x)$ avec $U$ suivant une loi uniforme standard et

$$Q(u|x) = u^{-\gamma(x)}[1 + \exp\{B_1^\top x + u^{-1}\}]^{-1},$$

où $B_1 = (0, 0, 5, 5, 0, \ldots, 0)^\top$.

Plusieurs modèles pour la fonction $\gamma(\cdot)$ sont considérés. Pour les deux premiers modèles, le sous-espace CTI est de dimension $q = 1$ admettant $B_0 = (2, 1, 0, \ldots, 0)^\top/3 \in \mathbb{R}^8$ pour base canonique.

- **Modèle 1** - Pour tout $x \in \mathcal{X}$,

$$\gamma(x) := \frac{1}{10} + \frac{9[\exp(2B_0^\top x) - 1]}{10[\exp(2) - 1]}.$$

- **Modèle 2** - Pour tout $x \in \mathcal{X}$,

$$\gamma(x) := \frac{1}{10} + \frac{9|\cos(3B_0^\top x)|}{10}.$$

Pour le modèle suivant le sous-espace CTI est de dimension $q = 2$ avec pour base canonique $B_0 = (e_1, e_2) \in \mathbb{R}^{8\times 2}$ où $e_1 = (1, 0, \cdots, 0)^\top$ et $e_2 = (0, 1, 0, \cdots, 0)^\top$.

- **Modèle 3** - Pour tout $x \in \mathcal{X}$,

$$\gamma(x) := \frac{1}{10} + \frac{9}{5}\left[(e_1^\top x - 0,5)^2 + (e_2^\top x - 0,5)^2\right].$$

Nous générons, pour chacun des modèles, $N = 100$ échantillons de taille $n = 2000$. Afin d'évaluer la performance d'un estimateur $\check{B}_n$ de la base $B_0$ du sous-espace CTI, on calcule la distance

$$\|\check{B}_n - B_0\|_F$$

où $\|\cdot\|_F$ est norme de Frobenius. Nous comparons l'estimateur TIDR proposé dans ce travail avec les estimateurs TIREX1 et TIREX2 proposés par Aghbalou et al. [1]. Notons toutefois que notre méthode n'est applicable que si nous avons pour tout $x \in \mathcal{X}$, $\gamma(x) > 0$ contrairement aux méthodes TIREX qui sont applicables quelque soit le signe de $\gamma(x)$. A noter également que la méthode TIREX1 n'est justifiée théoriquement que dans le cas $q = 1$.

Nous supposons pour le moment que la dimension $q$ du sous-espace CTI est connue. Une procédure de choix de $q$ sera proposée par la suite. Notre méthode nécessite de choisir deux paramètres $h$ et $\alpha$. Après plusieurs tests sur de nombreuses simulations, les choix $h = (n^{-1/3}/2^{q-1})^{1/q}$ et $\alpha = n^{-3/10}$ semblent donner de bons résultats dans la majorité des situations. Les méthodes TIREX dépendent elles d'un paramètre $k$ (nombre maximal des plus grandes observations utilisées pour l'estimation). La procédure choix par validation croisée

proposée par [1] est uniquement adaptée dans un contexte d'apprentissage supervisé. Pour donner à notre concurrent l'avantage maximale, nous prenons pour chaque simulation la valeur de $k$ donnant les meilleurs résultats parmi $\mathcal{K} := \{50, 60, \ldots, 390, 400\}$..

**Comparaison entre modèles** – Notons $\widehat{B}_n^{\mathrm{TIDR}}$, $\widehat{B}_n^{\mathrm{TIREX1}}$ et $\widehat{B}_n^{\mathrm{TIREX2}}$ les estimateurs de la base $B_0$ obtenus respectivement par les méthodes TIDR, TIREX1 et TIREX2. Les boîtes-à-moustaches des erreurs sont donnés dans les Figures 1 et 2. Dans tous les cas, la méthode TIDR donne les meilleurs résultats.

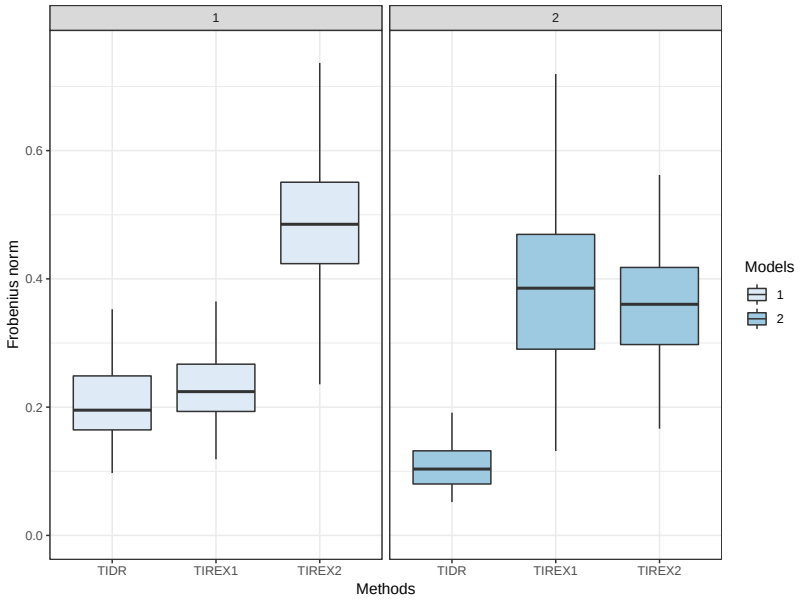



Figure 1: Comparaison des performances des méthodes TIDR et TIREX (1 et 2) sur un échantillon taille $n = 2000$, pour les modèles 1 et 2 avec $p = 8$ et $q = 1$.

**Estimation de l'indice des valeurs extrêmes** – Nous pouvons utiliser les estimateurs $\check{B}_n \in \{\widehat{B}_n^{\mathrm{TIDR}}, \widehat{B}_n^{\mathrm{TIREX1}}, \widehat{B}_n^{\mathrm{TIREX2}}\}$ du sous-espace CTI pour estimer l'indice des valeurs extrêmes conditionnel $\gamma(x)$ pour un point $x \in \mathcal{X}$. Nous choisissons pour ce faire d'utiliser l'estimateur local de Hill $\widehat{\xi}_B^{(H)}(B^\top x)$ introduit dans le paragraphe précédent.

Pour évaluer la qualité de l'estimation, nous calculons pour chaque réplication l'erreur moyenne quadratique

$$\mathrm{E}_\gamma(\check{B}_n) := \frac{1}{1000}\sum_{\ell=1}^{1000}\left[\widehat{\xi}^{(H)}_{\check{B}_n}(\check{B}_n^\top x_\ell) - \gamma(x_\ell)\right]^2,$$

où les 1000 points $x_\ell$ sont tirés uniformément sur $[0,1]^p$.

Les erreurs moyennes sur les $N = 100$ réplications sont rassemblées dans la Table 1. Comme on pouvait s'y attendre à la vue des résultats précédents, l'estimation de l'indice des valeurs extrêmes basé sur la réduction de dimension TIDR est meilleures pour les trois modèles considérés.

| Model | $\mathrm{E}_\gamma(\widehat{B}_n^{\mathrm{TIDR}})$ | $\mathrm{E}_\gamma(\widehat{B}_n^{\mathrm{TIREX1}})$ | $\mathrm{E}_\gamma(\widehat{B}_n^{\mathrm{TIREX2}})$ |
|---|---|---|---|
| 1 | **0.022** (0.018) | 0.023 (0.014) | 0.046 (0.028) |
| 2 | **0.042** (0.025) | 0.106 (0.043) | 0.095 (0.036) |
| 3 | **0.048** (0.016) | 0.194 (0.076) | 0.085 (0.022) |

Table 1: Comparaison des erreurs moyennes de l'estimation du tail-index (entre parenthèses l'écart type).

Dans la Table 2, nous avons également évalué la performance de l'estimateur local de Hill dans 2 cas : 1) on utilise la vraie base $B_0$ du sous-espace TIDR et 2) on n'effectue pas de réduction de dimension (i.e., on prend la matrice identité sur $\mathbb{R}^p$). Comme cela était prévisible l'erreur $\mathrm{E}_\gamma(Id)$ est importante justifiant ainsi la nécessité d'appliquer une procédure de réduction de dimension dans le cadre des valeurs extrêmes. Pour tous les modèles, les performances de notre méthode sont proches du cas où $B_0$ est connue.

| Model | $\mathrm{E}_\gamma(B_0)$ | $\mathrm{E}_\gamma(Id)$ |
|---|---|---|
| 1 | 0.014 (0.009) | 0.616 (0.102) |
| 2 | 0.035 (0.017) | 0.925 (0.135) |
| 3 | 0.047 (0.011) | 0.733 (0.109) |

Table 2: Comparaison des erreurs moyennes de l'estimation du tail-index (entre parenthèses l'écart type).

**Choix de la dimension du TIDR** – Nous avons jusqu'à présent supposer que la dimension l'espace CTI $q$ était connue. Évidement, ce n'est pas le cas en pratique. Nous proposons ici une procédure de choix basée sur le résultat suivant.

**Lemme 1.** *Soit $d \in \mathbb{N}\backslash\{0\}$. Nous avons*

| | Choice of dimension | | | $\mathrm{E}_\gamma(\widehat{B}_n^{\mathrm{TIDR}})$ | | |
|---|---|---|---|---|---|---|
| Model | $\widehat{q}_n = 1$ | $\widehat{q}_n = 2$ | $\widehat{q}_n = 3$ | $q = 1$ | $q = 2$ | $q = 3$ |
| 1 | 100 % | 0 % | 0 % | 0.024 | 0.041 | 0.064 |
| 2 | 100 % | 0 % | 0 % | 0.042 | 0.110 | 0.168 |
| 3 | 3 % | 97 % | 0 % | 0.074 | 0.048 | 0.095 |

Table 3: Résultat de la procédure de choix de la dimension et comparaison des erreurs moyennes selon le choix de $q$.

$$\min_{B \in \mathcal{B}_{d+1}} \Psi(B) \leq \min_{B \in \mathcal{B}_d} \Psi(B).$$

Posons pour tout $d$

$$c(d) := \min_{B \in \mathcal{B}_d} \Psi(B).$$

D'après le Lemme 1, les valeurs de $c(d)$ pour $d \in \{1, \ldots, q_0\}$ diminue quand $d$ augmente. Nous proposons donc de sélectionner la dimension en comparant successivement $\widehat{c}_n(d)$ à $\widehat{c}_n(d+1)$, où pour tout $d \in \{1, \ldots, p\}$,

$$\widehat{c}_n(d) := \widehat{\Psi}_n(\widehat{B}_n^{(d)}),$$

avec $\widehat{B}_n^{(d)}$ l'estimateur donné par la méthode TIDR pour la dimension $d$.
Plus précisément, nous choisissons

$$\widehat{q}_n = \min\{\ d \mid \widehat{c}_n(d) < \widehat{c}_n(d+1)\}.$$

Comme le montre la Table 3, cette procédure sélectionne la bonne dimension dans quasiment tous les cas. La qualité de l'estimation de $\gamma(x)$ est aussi comparée selon la dimension sélectionnée afin d'évaluer l'impact de ce choix.

# Références bibliographiques


[1] Aghbalou, A., F. Portier, A. Sabourin, and C. Zhou 2024. Tail inverse regression: dimension reduction for prediction of extremes. *Bernoulli* 30(1).

[2] Bellman, R. 1961. *Adaptive Control Process: A Guided Tour.* : Princeton University Press.

[3] Brazauskas, V., B. Jones, M. Puri, and R. Zitikis 2008. Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference* 138(11):3590–3604.

[4] Daouia, A., L. Gardes, and S. Girard 2013. On kernel smoothing for extremal quantile regression. *Bernoulli* 19(5B).

[5] Gardes, L. and G. Stupfler 2014. Estimation of the conditional tail index using a smoothed local hill estimator. *Extremes* 17:45–75.

[6] Goegebeur, Y., A. Guillou, and G. Stupfler 2015. Uniform asymptotic properties of a nonparametric regression estimator of conditional tails. *Annales de l'IHP Probabilités et statistiques* 51(3):1190–1213.

[7] Read, L. and R. Vogel 2015. Reliability, return periods, and risk under nonstationarity. *Water Resources Research* 51(8):6381–6398.

[8] Rockafellar, R. and S. Uryasev 2002. Conditional value-at-risk for general loss distributions. *Journal of banking & finance* 26(7):1443–1471.

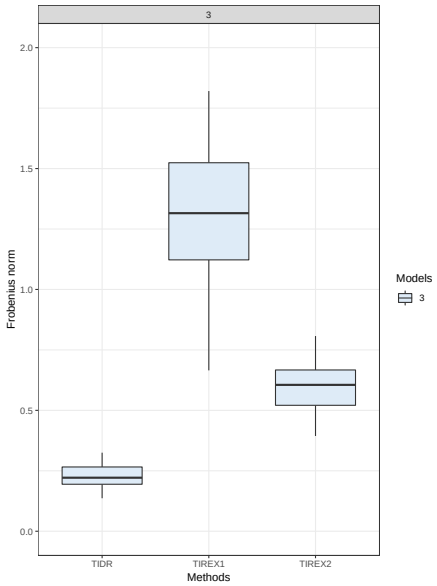



Figure 2: Comparaison des performances des méthodes TIDR et TIREX (1 et 2) avec un échantillon taille $n = 2000$, pour le modèle 3 avec $p = 8$ et $q = 2$.

# GEV-Extremal random forest

Lucien M. Vidagbandji[1] & Alexandre Berred [2] & Cyrille Bertelle [3] & Laurent Amanton [4]

[1] *Univ le Havre Normandie, LMAH, France, mahutin-lucien.vidagbandji@univ-lehavre.fr*
[2] *Univ le Havre Normandie, LMAH, France, alexandre.berred@univ-lehavre.fr*
[3] *Univ le Havre Normandie, LITIS, France, cyrille.bertelle@univ-lehavre.fr*
[4] *Univ le Havre Normandie, LITIS, France, laurent.amanton@univ-lehavre.fr*

**Résumé.** La régression quantile est une méthode statistique couramment utilisée dans l'analyse de régression. Contrairement à la régression classique, qui se concentre sur la prédiction de la moyenne conditionnelle d'une variable dépendante en fonction des variables indépendantes, la régression quantile vise plutôt à prédire les quantiles conditionnels. Les méthodes classiques de régression quantile rencontrent des défis, particulièrement lorsque le quantile d'intérêt est extrême, en raison du nombre limité de données disponibles dans la queue de la distribution, ou lorsque la fonction quantile est complexe. Dans le cadre de cette étude, nous proposons une méthode de régression quantile extrême basée sur la théorie des valeurs extrêmes et l'apprentissage statistique pour surmonter ces défis. Conformément à l'approche de maxima de bloc (BM) de la théorie des valeurs extrêmes, nous approchons la distribution conditionnelle des BM par la distribution des valeurs extrêmes généralisée, dont les paramètres dépendent des covariables. Pour estimer ces paramètres, nous utilisons une méthode basée sur les forêts aléatoires généralisées. Les résultats obtenus à partir d'applications sur des données simulées mettent en évidence que notre méthode est compétitive avec d'autres approches de régression quantile.

**Mots-clés.** Régression quantile, Distribution des valeurs extrême généralisée, Forêt aléatoire généralisée, Maximum de vraisemblance, Bloc maxima, quantile extrême.

**Abstract.** Quantile regression is a commonly used statistical method in regression analysis. Unlike classical regression, which focuses on predicting the conditional mean of a dependent variable based on independent variables, quantile regression aims to predict conditional quantiles. Classical quantile regression methods face challenges, especially when the quantile of interest is extreme, due to the limited number of data available in the tail of the distribution or when the quantile function is complex. In this study, we propose an extreme quantile regression method based on extreme value theory and statistical learning to overcome these challenges. Following the Block Maxima (BM) approach of extreme value theory, we approximate the conditional distribution of BM by the generalized extreme value distribution, with parameters depending on covariates. To estimate these parameters, we employ a method based on generalized random forests. Results obtained from applications on simulated data highlight that our method is competitive with other quantile regression approaches.

**Keywords.** Quantile regression, Generalized Extreme Value Distribution, Generalized Random Forest, Maximum Likelihood, Block Maxima, Extreme Quantile.

# 1 Introduction

The modeling of extreme phenomena is crucial in various fields such as finance, meteorology, public health, and many others. Understanding the tails of the distribution of random variables is essential for assessing risks associated with rare but potentially devastating events. In this context, Extreme Value Theory (EVT) has emerged as a powerful tool for characterizing the behavior of extremes in a distribution. The main objective of this work is to explore extreme quantile regression by combining the fundamental principles of EVT with statistical learning techniques. Extreme quantile regression stands out for its ability to specifically model the quantiles of the distribution, providing detailed information about extremes. This approach is particularly valuable in contexts where a focus on extreme values is crucial for making informed decisions. Specifically, if $Y \in \mathcal{Y} \subset \mathbb{R}$ represents a random variable describing a risk factor dependent on a set of covariates represented by the random vector $X \in \mathcal{X} \subset \mathbb{R}^p$, the goal is to estimate the conditional extreme quantile given by :

$$\mathcal{Q}_\tau(x) = \inf\{y : F^{-1}_{Y|X=x}(y) \geq \tau\} \tag{1}$$

with $\tau$ close to 1 and $F^{-1}_{Y|X=x}$ the generalized inverse of the conditional distribution $Y|X = x$ (Koenker *et al.* (1978)).

Letting $n$ be the sample size available for analysis and $\tau = \tau_n$ (depending on the sample size) be the order of the quantile we seek to estimate, classical methods for quantile estimation work well when $n(1 - \tau_n) \to \infty$ as $\tau_n \to 1$ (as $n \to +\infty$). In this case, the quantile to be estimated is within the sample, and there is also a large amount of data in the quantile range to be estimated. However, the situation is different when $n(1-\tau_n) \in [0, +\infty[$ and $\tau_n \to 1$ (as $n \to +\infty$). In the latter case, estimation requires extrapolation beyond the data range or into the tail of the distribution. In other words, the sought-after quantile is outside the range of the available sample, making estimation more complex and requiring specific approaches to handle these boundary situations. Thanks to the asymptotic results of extreme value theory (de Haan *et al.* (2006)), extrapolation beyond this data range is possible. Quantile regression methods face other challenges, including the complexity of the quantile function or the high dimensionality of the feature vector. To address the latter, we will use statistical learning methods, primarily the generalized random forests method by Athey *et al.* (2019), which is an extension of Breiman's classical random forests (2001).

Several works are proposed to address these different challenges of quantile regression. To address the first challenge, methods based on extreme value theory have been developed (see Chernozhukov *et al.* (2017) for an overview), while for the second, approaches based on statistical learning have been developed (*Meinshausen et al.* (2006), *Athey et al.* (2019), etc.). Recently, methods have been proposed that combine the Peaks Over Threshold (POT) approach of extreme value theory with statistical learning methods (Youngman (2018), *Farkas et al.* (2021), *Velthoen et al.* (2023), *Pasche et al.* (2023), *Gnecco et al.* (2024)). Our work aims to adapt the method of *Gnecco et al.* (2024) to the block maxima approach of extreme value theory. Explicitly, we model the conditional distribution of the relationship (1) using the generalized extreme value distribution, with parameters varying depending on the feature vector. These parameters are estimated by minimizing a local likelihood, weighted by weights obtained using generalized random forests.

# 2 Background

## 2.1 Block Maxima Approach

In this section, we will review some concepts regarding the approach of Extreme Value Theory (EVT), which we will use to address the first challenge of quantile regression stated in the introduction. The Block Maxima (BM) method is based on the limiting distribution of the maximum of a sequence of random variables $X_1, \cdots, X_m$ drawn independently and identically from a variable $X$ with probability distribution $F$, as shown by Fisher *et al.* (1928), Gnedenko *et al.* (1943). These authors demonstrated that there exist normalization constants $a_m > 0$ and $b_m \in \mathbb{R}$ such that

$$\lim_{m \to +\infty} F^m(a_m x + b_m) = G_\xi(x), \ x \in \mathbb{R} \tag{2}$$

where $G_\xi$ is a non-degenerate probability distribution defined by :

$$G_\xi(x) = \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right),$$

with $1 + \xi x > 0$. Any function $F$ satisfying equation (2) belongs to the max-domain of attraction of the distribution of extreme values $G_\xi$ and is denoted as $F \in \mathcal{D}(G_\xi)$ (De Haan *et al.* (2006)).

If we consider $Y_1, \cdots, Y_N$ as a sequence of independent and identically distributed random variables according to the random variable $Y$ with cumulative distribution function $F \in \mathcal{D}(G_{\xi_0})$ and corresponding normalization constants $a_n$ and $b_n$, the Block Maxima (BM) method involves dividing the data into $n$ blocks of the same size $m > 1$ (or nearly the same), denoted as $B_{k,m} = \{Y_{(k-1)m+1}, \cdots, Y_{km}\}$, where $k = 1, \cdots, n$. For any $m > 1$, the distribution of $Z_k = \max_{B_{k,m}}(Y_i)$ is $F^m$ and satisfies (2), thus its distribution is approximated by the generalized extreme value (GEV) distribution with parameters $(a_m, b_m, \xi_0)$. The BM method assumes that the distribution of these block maxima, denoted as $Z_k$, exactly follows the GEV distribution, and the resulting sequence of random variables $Z_1, \cdots, Z_n$ is also independent and identically distributed. Note that the choice of block size is crucial : increasing the block size results in large estimation variance and decreasing it will introduce bias in estimation. This boils down to a trade-off between bias and variance. For accurate estimation, a trade-off between bias and variance must be found when defining the blocks. The BM method is presented and discussed in the literature, and notable references include the books by Coles (2001) and De Haan *et al.* (2006). The GEV distribution is given by :

$$G_{\mu,\sigma,\xi}(z) = \begin{cases} \exp\left(-(1+\xi\frac{z-\mu}{\sigma})_+^{-\frac{1}{\xi}}\right) & \xi \neq 0 \\ \exp\left(-\exp(-\frac{z-\mu}{\sigma})\right) & \xi = 0 \end{cases}, \ \forall z \in \mathbb{R} \tag{3}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$ are the parameters of location, scale, and shape, respectively, and $a_+ = \max\{0, a\}$. The quantile of order $\tau$ of GEV is obtained through equation (1) and is given by :

$$Q_\tau = \begin{cases} \mu + \frac{\sigma}{\xi}\left((-\ln(\tau))^{-\xi} - 1\right), & \text{if } \xi \neq 0 \\ \mu + \xi \ln\left(-\ln\left(\tau\right)\right), & \text{if } \xi = 0 \end{cases} \tag{4}$$

Estimating the quantile of GEV amounts to estimating the parameters $\mu$, $\sigma$, and $\xi$. There are several methods for estimating these parameters, with the most common being the maximum likelihood method. Our proposed method is based on a variation of this estimator.

## 2.2 Generalized random forests

The Generalized Random Forest (GRF) is an extension of the classical Random Forest method proposed by Breiman (2001). The Random Forest is an ensemble method for regression and classification that aggregates $B$ trees fitted in parallel on bootstrap samples from the training dataset. Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be the real response variable and $X \in \mathcal{X} \subset \mathbb{R}^p$ be the vector of covariates. Classical regression analysis aims to obtain an estimate of the conditional mean $\eta(x) = \mathbb{E}(Y|X = x)$ of the response variable $Y$, given $X = x$. This is achieved by minimizing the expected quadratic loss :

$$\eta(x) = \arg\min_{z \in \mathcal{Y}} \mathbb{E}\left(l(Y - z)|X = x\right) \text{ with } l(y_1, y_2) = (y_1 - y_2)^2 \tag{5}$$

If $\eta_b(x)$ is the value predicted by the $b$-th tree for the test data $x \in \mathbb{R}^p$, in the case of regression, it is given by :

$$\hat{\eta}_b(x) = \sum_{i=1}^{n} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|} Y_i, \ b = 1, \cdots, B$$

where $R_b(x) \in \mathbb{R}^p$ denotes the region containing $x$ in tree $b$ and $|E|$ is the cardinality of set $E$. The prediction made by the Random Forest is given by :

$$\begin{aligned} \hat{\eta}(x) &= \frac{1}{B} \sum_{b=1}^{B} \hat{\eta}_b(x) \\ &= \sum_{i=1}^{n} w_n(x, X_i) Y_i \end{aligned} \tag{6}$$

with

$$w_n(x, X_i) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|}. \tag{7}$$

Each $w_n(x, X_i)$ in the Random Forest represents a similarity weight.

The Generalized Random Forest (GRF), introduced by Athey et al. in 2019, represents an extension of the Random Forest method. This approach preserves the attractive features of classical Random Forests while providing the flexibility to use custom loss functions for tree construction in the forest, i.e., the function $l$ used in (5). An additional advantage of the Generalized Random Forest is that the similarity weights it generates capture the heterogeneity of the quantile function, unlike classical Random Forests. Indeed, the similarity weight $w_n(x, X_i)$ estimated by the classical forest is high for an observation $X_i$ when $\mathbb{E}(Y|X = X_i) \approx \mathbb{E}(Y|X = x)$, but there are situations where this weight is high and $\mathcal{Q}(Y|X = X_i) \not\approx \mathcal{Q}(Y|X = x)$ (Athey *et al.* (2019), Gnecco*et al.* (2024)). We use the

Generalized Random Forest in our method to obtain the necessary similarity weights for estimating the parameters of the conditional GEV distribution. These parameters are then used to estimate the conditional quantile, as explained in Section (3).

### 2.3 Quantile Regression

When the conditional distribution $F(.|X = x)$ is continuous, the conditional quantile function given in equation (1) simplifies to :

$$\mathcal{Q}_\tau(x) = F^{-1}(\tau|X = x). \tag{8}$$

To our knowledge, the first appearance of random forest methods in the context of quantile regression dates back to Meinshausen (2006). He estimates the conditional distribution as follows :

$$\hat{F}\left(y|X = x\right) = \sum_{i=1}^{n} w_n(x, X_i)\mathbb{1}_{\{Y_i \leq y\}},$$

where the weights $w_n(x, X_i)$ are obtained from the classic random forest as defined in (7), and this estimated distribution is then used in (1) to obtain the conditional quantile. Another method is that of Athey *et al.* (2019), which is an application of generalized random forests with the quantile loss function given by $\rho_\tau(c) = c(\tau - \mathbb{1}_{\{c<0\}})$. In this work, we will use the term generalized random forest (grf) to denote the forest built using the quantile loss. Other methods are proposed to address the challenges outlined in the introduction by combining the peaks-over-threshold (POT) approach from extreme value theory with machine learning. Notable works include *Farkas et al.* (2021) using regression trees, *Velthoen et al.* (2023) with gradient boosting, *Pasche et al.* (2023) using neural networks, and *Gnecco et al.* (2024) employing generalized random forests. Our method is an extension of the BM approach proposed by *Gnecco et al.* (2024) and is explained in the following section.

## 3 GEV Extremal Random Forest

We propose a method for extreme conditional quantile regression addressing the challenges outlined in the introduction. To address the first issue, we model the tail of the conditional distribution $F(.|X = x)$ in equation (8) with the generalized extreme value distribution. To address the second issue, we use generalized random forests with quantile loss as the loss function to obtain the weights $w_n(x, X_i)$, which we use to estimate the parameters of the GEV distribution as weighted likelihood estimators, as proposed by Gnecco *et al.* (2024) to solve similar problems in the peaks-over-threshold (POT) approach.

In the conditional case, the parameters of the distribution GEV are functions of the covariate vector $x$. The conditional GEV distribution is obtained by replacing $\theta = (\mu, \sigma, \xi)$ in (3) with $\theta(x) = (\mu(x), \sigma(x), \xi(x))$ (where $\mu, \sigma, \xi : \mathcal{X} \to \mathbb{R}$). The quantile of order $\tau$ (with

$\tau$ close to 1) of conditional GEV is obtained through equation (8) and is given by :

$$Q_\tau(x) = \begin{cases} \mu(x) + \dfrac{\sigma(x)}{\xi(x)} \left( (-\ln(\tau))^{-\xi(x)} - 1 \right), & \text{if } \xi(x) \neq 0 \\ \mu(x) + \xi(x) \ln\left(-\ln\left(\tau\right)\right), & \text{if } \xi(x) = 0 \end{cases} \tag{9}$$

Estimating the conditional GEV quantile amounts to estimating the parameters $\mu(x)$, $\sigma(x)$, and $\xi(x)$. Our proposed method involves estimating these parameters and then substituting them into equation (9) to obtain an estimation of $Q_\tau(x)$.

Here, we propose an alternative form of the classical maximum likelihood estimator of the GEV distribution. As suggested by Gnecco *et al.* (2024) for the POT approach, we estimate $\theta(x)$ by $\hat{\theta}(x) = (\hat{\mu}(x), \hat{\sigma}(x), \hat{\xi}(x))$, which is the weighted maximum likelihood estimator, i.e., minimizing

$$L_n(\theta, x) = \sum_{i=1}^{n} w_n(x, X_i) l_{\theta(x)}(z_i) \tag{10}$$

with $l_\theta(z_i)$ such that :

— when $\xi(x) \neq 0$,

$$l_{\theta(x)}(z_i) = \log(\sigma(x)) + (1 + \frac{1}{\xi(x)}) \log\left(1 + \xi(x)\frac{z_i - \mu(x)}{\sigma(x)}\right) + \left(1 + \xi(x)\frac{z_i - \mu(x)}{\sigma(x)}\right)^{\frac{-1}{\xi(x)}}$$

if $1 + \xi(x)\frac{z_i - \mu(x)}{\sigma(x)} > 0$ and $+\infty$ otherwise.

— when $\xi(x) = 0$,

$$l_{\theta(x)}(z_i) = \log(\sigma(x)) + \left(\frac{z_i - \mu(x)}{\sigma(x)}\right) + \exp\left(\frac{z_i - \mu(x)}{\sigma(x)}\right)$$

the $w_n(x, X_i, )$ $i = 1, \cdots, n$ are obtained using the generalized random forests method. The weighted maximum likelihood estimator is defined as :

$$\hat{\theta}(x) = \arg \min_{\theta(x) \in \Theta} L_n(\theta, x).$$

The parameter $\xi$ is important as it determines the shape of the distribution's tail. Therefore, significant attention is paid to it in the estimation of the GEV distribution in the literature. For instance, Bücher *et al.* (2020) propose a maximum likelihood estimator for the GEV distribution by penalizing only the parameter $\xi$, in the non-conditional case. Gnecco *et al.* (2024) also suggested a maximum likelihood estimator for the generalized Pareto distribution by penalizing the parameter $\xi$ in the conditional case. Building upon these works, primarily on the study by Gnecco *et al.* (2024), we also penalize $\xi$, considering the penalized maximum likelihood estimator defined as :

$$\hat{\theta}_{\text{pena}}(x) = \arg \min_{\theta(x) \in \Theta} L_n(\theta, x) + \lambda(\xi - \hat{\xi})^2, \tag{11}$$

where $0 \leq \lambda$ is the penalty parameter, and $\hat{\xi}$ is considered as the shape parameter of the maximum likelihood estimator obtained according to equation (10) with weights $w_n(x, X_i)$ all equal to 1, as used in Gnecco *et al.* (2024) for the POT approach. Cross-validation method is employed for the selection of $\lambda$.

# 4 Simulation results

This section presents some application results of our method on simulated data, based on the simulation part of the works by Gnecco *et al.* (2024) and Velthoen *et al.* (2023). We generate $N = 90000$ (denoted *ntrain* in the figures) i.i.d. samples of $X \sim \mathcal{U}_{[-1,1]^p}$ and the conditional distribution $Y|X = x \sim \gamma(x)T_{\nu(x)}$, where $T_k$ is the Student's distribution with $k$ degrees of freedom. For the figures below, we use $\gamma(x) = 1 + \mathbb{1}_{\{x_1>0\}}$ (denoted *model1* in the figures) and $\nu(x) = 4 - (x_1^2 - 2x_2^2 + x_3^2)$ (denoted *quadratic* in the figures) for any test data $x \in \mathbb{R}^p$, where $x_i$ denotes the $i^{th}$ component of $x$, and $p = 40$. The conditional quantile function depends only on the first three components of $x$, and the remaining components are noise. Our method involves dividing the $N$ data into $n$ blocks of size 30 and considering the maxima of the formed blocks, so only $n = 3000$ data are considered for the training process to obtain the weights $w_n(x, X_i)$ and the adjustment of the likelihood given in (11). We evaluate our method on test data $\{x^i\}_{i=1}^{N'}$ with $N' = 3000$ (denoted $n_{\text{test}}$ in the figures), independent of the training data and generated by the Halton sequence on the cube $[-1, 1]^p$ (with $x^i \in \mathbb{R}^p$ representing the $i^{th}$ data of the test sample), where we divide into blocks of size 30 and consider only the maxima per block as test data. Thus, only $n' = 100$ data are considered for the model evaluation.

To highlight the performance of our method, designated as GEV, on the graphs below, we compare its Mean Integrated Squared Error (MISE) with other approaches using statistical learning. This includes the **quantile regression forests** by Meinshausen (2006), denoted as QRF, the method of **generalized random forests** by Athey *et al.* (2019), denoted as GRF, and the unconditional method denoted as Uncond. Note that all the models considered are trained on the same dataset, which consists of the formed maxima, hence a set of $n = 3000$ training data. The test is also performed on the same dataset, namely a set of $n' = 100$ test data.

The results show that our method performs better for conditional quantile estimation, mainly for quantiles close to 1, more clearly for the quantile order from $\tau = 0.99$ to $\tau = 0.9999$, as shown in Figure (1). Figure (2) shows the variation of MISE as a function of the predictor size $p$ for a quantile order fixed at $\tau = 0.999$, where we see that our method is competitive, mainly with the GRF method. We also see that it performs better than other models for $p = 40$.

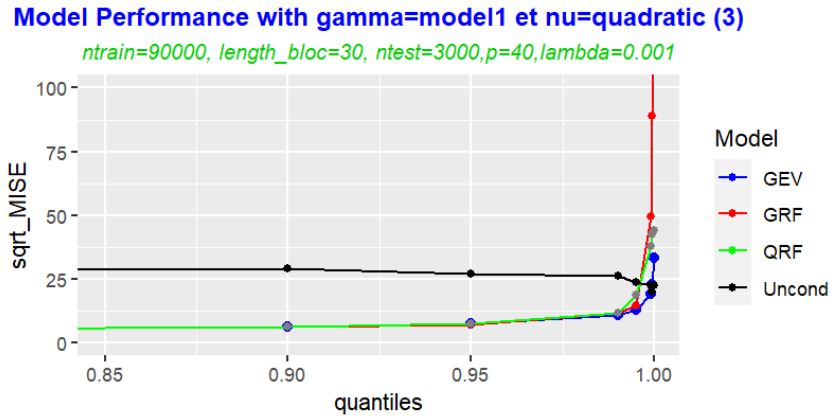



FIGURE 1 – Experiment 1

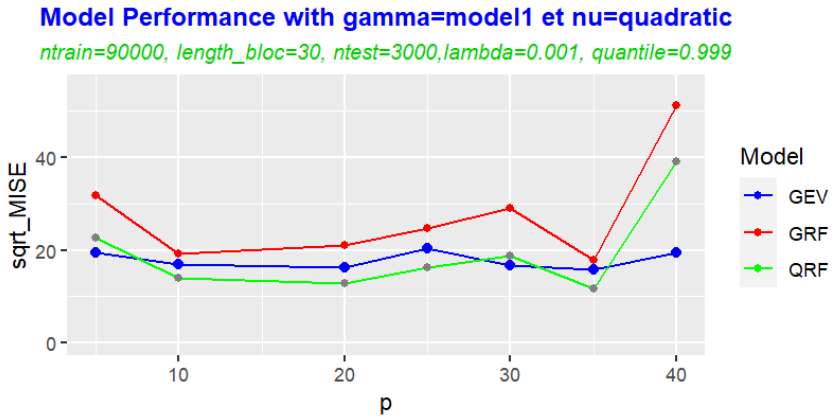



FIGURE 2 – Experiment 1

# Bibliography

Coles, S., Bawa, J.,Trenner, L. et al. (2001), An introduction to statistical modeling of extreme values, *Springer*, 208.

Bücher, A.,Lilienthal,J. et al., Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis, *Extremes*, 24, pp 325-348.

Meinshausen, N. (2006), Quantile Regression Forests, *Journal of Machine Learning Research*, 7, pp. 983-999.

Breiman, L. (2001), Random forests, *Machine Learning*, 45, pp 5–32.

Athey, S. and Tibshirani, J. and Wager, S. (2019), Generalized random forests, *The Annals of Statistics*, 2, pp 1148-1178.

Gnecco, N., Terefe, E. M. and Engelke, S. (2024), Extremal Random Forests, *Journal of the American Statistical Association*, 0, pp 1-24.

Velthoen, J., Dombry, C., Cai, JJ. et al. (2023), Gradient boosting for extreme quantile regression, Extremes, 26, pp 639–667.

Pasche, O. C. and Engelke, S.(2023), Neural Networks for Extreme Quantile Regression with an Application to Forecasting of Flood Risk, arXiv.

Dombry, C. (2015), Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework, *Bernoulli*, 1, pp 420–436.

Koenker, R., Bassett G. (1978), Regression Quantiles, *Econometrica*, 46(1) :33.

Chernozhukov, V., Fernández-Val, I., Kaji, T. (2016), Extremal quantile regression : An overview, *arXiv preprint arXiv :1612.06850*

Youngman, B. (2019), Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts, *Journal of the American Statistical Association*, 114, pp 1865–1879.

# A de-randomization argument for estimating extreme value parameters of heavy tails

Abdelaati Daouia[1] & Joseph Hachem[2] & Gilles Stupfler[3]

[1] *Université Toulouse Capitole, TSE, 1 Esplanade de l'Université, 31000 Toulouse, France,* abdelaati.daouia@tse-fr.eu

[2] *Université Toulouse Capitole, TSE, 1 Esplanade de l'Université, 31000 Toulouse, France,* joseph.hachem@tse-fr.eu

[3] *Université d'Angers, CNRS, LAREMA, SFR MATHSTIC, F-49000 Angers, France,* gilles.stupfler@univ-angers.fr

**Résumé.** En analyse des valeurs extrêmes, il a été récemment montré qu'on peut utiliser une technique de dé-randomisation, consistant à remplacer un seuil aléatoire dans l'estimateur d'intérêt par son homologue déterministe, afin d'estimer simultanément plusieurs risques extrêmes, mais seulement pour des données i.i.d.. Dans cet exposé, nous montrerons comment cette méthode peut être utilisée pour estimer plusieurs quantités extrêmes (indice de queue, expected shortfalls...) dans des contextes généraux de données dépendantes/hétéroscédastiques/hétérogènes, sous une hypothèse $L^1$ pondérée sur l'écart entre la loi moyenne des données et la loi dominante. Cette technique peut également être utilisée pour traiter des données hétérogènes multivariées, ce que la littérature actuelle ne permet pas de faire.

**Mots-clés.** Valeurs extrêmes, dé-randomisation, modèles à queue lourde, estimateur de Hill, données hétérogènes, statistiques d'ordre.

**Abstract.** In extreme value analysis, it has recently been shown that one can use a de-randomization trick, replacing a random threshold in the estimator of interest with its deterministic counterpart, in order to estimate several extreme risks simultaneously, but only in an i.i.d. context. In this talk, I will show how this method can be used to handle the estimation of several tail quantities (tail index, expected shortfall...) in general dependence/heteroskedasticity/heterogeneity settings, under a weighted $L^1$ assumption on the discrepancy between the average distribution of the data and the prevailing distribution. This technique can also be used to deal with multivariate heterogeneous data, which cannot be handled with current methods.

**Keywords.** Extreme values, de-randomization, heavy tails, Hill estimator, heterogeneity, order statistics.

## Summary of the presentation

Let $n \geq 1$, $X_1^{(n)}, \ldots, X_n^{(n)}$ be (almost surely finite) random variables, and denote by $X_{1:n}^{(n)} \leq X_{2:n}^{(n)} \leq \cdots \leq X_{n:n}^{(n)}$ their order statistics. The original motivation for this work is the analysis

of the asymptotic behavior of the quantity

$$\widehat{e}_{f,n}(k) = \frac{1}{k}\sum_{i=1}^{k} f(X^{(n)}_{n-i+1:n}) - f(X^{(n)}_{n-k:n})$$

with $k(n) = k \to \infty$ such that $k/n \to 0$. In the special situation when the $X_i^{(n)}$ have the same distribution as a random variable $X$ having quantile function $q$, the quantity $\widehat{e}_{f,n}(k)$ is a natural estimator of $\mathbb{E}(f(X) - f(q(1-k/n)) \,|\, X > q(1-k/n))$. This quantity is the mean excess value of $f(X)$ when $X > q(1-k/n)$, which motivates the name *mean $f$-excess*, and we will call $\widehat{e}_{f,n}(k)$ the *empirical mean $f$-excess* above the order statistic $X^{(n)}_{n-k:n}$ throughout. Prominent among these quantities is the one obtained with $f = \log$,

$$\widehat{e}_{\log,n}(k) = \frac{1}{k}\sum_{i=1}^{k} \log X^{(n)}_{n-i+1:n} - \log X^{(n)}_{n-k:n},$$

which is reminiscent of the Hill [1975] estimator of a positive extreme value index for heavy-tailed distributions. The behavior of the empirical mean $f$-excess is hard to assess in a non-i.i.d. scenario due to its formulation in terms of order statistics. It is however well understood in the i.i.d. case, as seen in Stupfler [2019], using a de-randomization technique to write $\widehat{e}_{f,n}(k)$ as the sum of a quotient of random sums of i.i.d. terms and a negligible remainder; the proof involves a Gaussian approximation of the underlying empirical process $\widehat{F}_n$. A recent advance has been made in Einmahl and He [2023] for the Hill estimator for independent heterogeneous data, using the empirical process theory. The main difficulty of this approach is that, on one hand, it does not allow for deriving the asymptotic bias of the Hill estimator and, on the other hand, it cannot be easily adapted to multivariate or dependent data, and that the assumptions made require some sort of uniform boundedness of the mean survival function of the sample, restricting the scope of such a technique.
To address these issues in a general framework, we shall adapt the de-randomization technique of Stupfler [2019]. First, we will show, under the hypothesis that there exist a positive sequence $x_n$ (which will typically be a quantile) tending to infinity and constants $c_1, c_2 > 0$ and $c_3 \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \frac{\frac{1}{k}\sum_{i=1}^{n} \mathbb{E}((f(X_i^{(n)}) - f(x_n))\mathbb{1}\{X_i^{(n)} > x_n\})}{x_n f'(x_n)} = c_1 \tag{1}$$

$$\text{and } \forall t \in \mathbb{R},\ \lim_{n\to\infty} \sqrt{k}\left(\frac{k/n}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X_i^{(n)} > (1+t/\sqrt{k})x_n)} - 1\right) = c_2 t + c_3, \tag{2}$$

that the asymptotic behavior of the empirical mean $f$-excess is very closely linked to the joint asymptotic behavior of its de-randomized counterpart

$$\overline{e}_{f,n}(k) = \frac{1}{k}\sum_{i=1}^{n} (f(X_i^{(n)}) - f(x_n))\mathbb{1}\{X_i^{(n)} > x_n\},$$

and, for $t \in \mathbb{R}$ fixed,

$$\widehat{\overline{F}}_n((1+t/\sqrt{k})x_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i^{(n)} > (1+t/\sqrt{k})x_n\},$$

which are way easier to handle in a general scenario, including the multivariate or dependent cases, which cannot be handled with current techniques; more specifically, (1) means that $\mathbb{E}(\overline{e}_{f,n}(k))$ has an appropriate asymptotic behavior, and (2) means that $X^{(n)}_{n-k:n,n}$ can basically be replaced with $\widehat{F}_n((1+t/\sqrt{k})x_n)$ for any $t \in \mathbb{R}$, using an argument inspired by Koenker [2005]. Then, we will show that these two estimators are jointly asymptotically normal in a broad heterogeneous scenario for independent data with simple hypotheses, assuming that there exist a distribution $\mathbb{P}_X$ of a heavy-tailed random variable $X$ and $\gamma > 0$ such that, for any sequence $(x_n)$ tending to infinity with $n\mathbb{P}(X > x_n) \to \infty$ (for example, $x_n = q(1-k/n)$ with $q$ the quantile function of $X$), and for all $u > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{P}(X_i^{(n)} > x_n u)}{\mathbb{P}(X > x_n)} = u^{-1/\gamma}.$$

The proof of the joint asymptotic normality of the above two estimators relies on elementary tools of probability theory such as the Lyapounov CLT and the Cramér-Wold device. It also uses elementary techniques in extreme value analysis such as the Potter bounds and Drees inequality, as seen in de Haan and Ferreira [2006], without resorting to Gaussian approximations. As a result, this approach can be used for multivariate data, and could be adapted for dependent data as well.

# Acknowledgments

This research was supported by the French National Research Agency under the grants ANR-19-CE40-0013 (ExtremReg project), ANR-18-EURE-0023 (EUR MINT), ANR-17-EURE-0010 (EUR CHESS) and ANR-11-LABX-0020-01 (Centre Henri Lebesgue). A. Daouia and G. Stupfler acknowledge financial support from the TSE-HEC ACPR Chair.

# References

L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, 2006.

J. H. J. Einmahl and Y. He. Extreme value inference for heterogeneous power law data. *Annals of Statistics*, to appear, 2023.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174, 1975.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

G. Stupfler. On a relationship between randomly and non-randomly thresholded empirical average excesses for heavy tails. *Extremes*, 22(4):749–769, 2019.

# Modeling moderate and extreme urban rainfall at high spatio-temporal resolution

Chloé Serre-Combe [1] & Nicolas Meyer [2] & Thomas Opitz [3] & Gwladys Toulemonde [4]

[1] *Univ. Montpellier, CNRS, IMAG, Inria, France, chloe.serre-combe@umontpellier.fr*
[2] *Univ. Montpellier, CNRS, IMAG, Inria, France, nicolas.meyer@umontpellier.fr*
[3] *INRAE, BioSP, Avignon, France, thomas.opitz@inrae.fr*
[4] *Univ. Montpellier, CNRS, IMAG, Inria, France, gwladys.toulemonde@umontpellier.fr*

**Résumé.** La modélisation des précipitations est d'un grand intérêt pour l'analyse des risques d'inondation. Nous proposons de modéliser la distribution des précipitations urbaines mesurées à haute résolution spatiale et temporelle par le réseau de pluviomètres de l'Observatoire Urbain de Montpellier sur quatre années de mesures. Nous les combinons avec des données de réanalyse radar pour étendre notre analyse à une période plus longue avec une résolution moins fine. Pour cette modélisation, nous considérons simultanément les pluies modérées et intenses en utilisant l'*Extended Generalized Pareto Distribution (EGPD)* pour éviter la sélection d'un seuil, souvent délicat en statistiques des extrêmes, mais aussi pour réduire la complexité de l'estimation des paramètres. Nous modélisons également la dépendance spatio-temporelle en incorporant l'advection à travers un processus spatio-temporel de Brown-Resnick. Nous utilisons des indices d'autocorrélation extrême, pour montrer sa variabilité entre les différents sites et à différents temps des mesures. Nous mettrons en évidence l'importance de considérer l'advection en comparant le modèle obtenu à un modèle séparable plus simple.

**Mots-clés.** Théorie des valeurs extrêmes, modélisation de précipitation, EGPD, haute résolution spatio-temporelle, modélisation de la dépendance, processus de Brown-Resnick, advection

**Abstract.** Precipitation modeling is of great interest for flood risk analysis. We propose to model the distribution of urban precipitation measured at high spatial and temporal resolution by the Montpellier Urban Observatory rain gauge network over four years of measurements. We combine them with radar reanalysis data to extend our analysis to a longer period with less fine resolution. For our modeling approach, we simultaneously consider moderate and intense rainfall by using the Extended Generalised Pareto Distribution (EGPD) to avoid explicit threshold selection, often tricky in extreme statistics, and to reduce the complexity of parameter estimation. We also model the spatio-temporal dependence by incorporating advection through a spatio-temporal Brown-Resnick process. We use indices of extreme autocorrelation, to show its variability between locations in relation to their spatial distances and to the temporality of the measurements. We will highlight the importance of including advection by comparing it with a simpler separable model.

**Keywords.** Extreme value theory, rainfall modeling, EGPD, high spatio-temporal resolution, dependence modeling, Brown-Resnick process, advection

# 1 Study

Managing flood risk requires a precise understanding of precipitation patterns. Severe flooding can indeed occur after extreme rainfall events, but also after moderate ones, for example if the latter last a long time or if the ground is already saturated with water. The risk is even greater in urban areas where water absorption is reduced by impervious surfaces. This shows the importance of understanding the behaviour of both extreme and moderate rainfall events and their spatio-temporal variability.

We focused our work on a specific area in Montpellier, south of France. Montpellier is known to be frequently exposed to significant rainfall events known as Mediterranean episodes, especially during the autumn season. This type of event is characterised by a large amount of rainfall in a short period of time and is very localised, causing local urban flooding. This encourages us to build a model which takes into account this high spatio-temporal resolution. Our study relies on 17 rain gauges in Montpellier in the water catchment of the Verdanson, a tributary of the Lez (see Figure 1). The rainfall measurements from these stations are provided by the urban Observatory of the HydroScience Montpellier (OHSM) (see Finaud-Guyot et al. (2023)). They cover the period from 2019 to 2022 and are given with a high temporal resolution, recorded minute by minute. In order to minimise measurement errors and to avoid very strongly discretized values, the data are aggregated in 5-minute intervals, which allows us to maintain a high temporal resolution. This fine resolution implies that we obtain a very small proportion of non-zero values: they represent only 1.2% of the data. In terms of spatial granularity, we also have a high resolution with a distance between two stations ranging from 77 to 1531 meters. We want to model the spatio-temporal characteristics of these rainfall data to better understand the rainfall behaviour over this area.

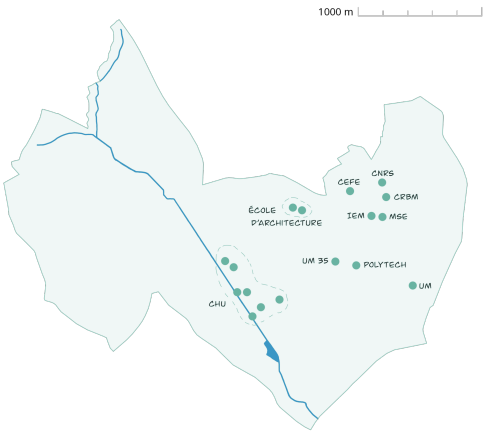


Figure 1: Rain gauges location

In order to refine our modeling, we consider another dataset: the French COMEPHORE[1] Mosaic from Météo France (see Tabary et al. (2012)). These reanalysis data, which combine radar and rain gauge measurements over France, have a lower resolution than the previous OHSM data. The rainfall measurements are provided with a spatial resolution of 1 km$^2$ per pixel. Each pixel represents the cumulative rainfall over one hour, and the time period ranges from 1997 to 2021. We extract measurements pixels on the Montpellier area.

**Notations:** Let $\mathcal{S} \subset \mathbb{R}^2$ be our spatial domain and let $\mathcal{T} \subset \mathbb{R}_+$ be our temporal domain with equidistant time points.

[1]COmbinaison en vue de la Meilleure Estimation de la Précipitation HOraiRE, dataset is available on AERIS platform (https://radarsmf.aeris-data.fr/)

## 2 Univariate modeling

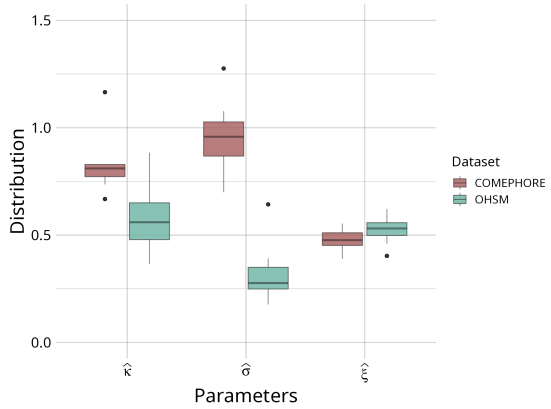
In this section, we focus on characterizing the behaviour of moderate and extreme precipitation at individual locations. In a univariate framework, let $X_s$ be the random variable representing the amount of rainfall in millimeters measured at a given location $s \in \mathcal{S}$.

In order to take the full range of precipitation into account, we use the Extended Generalized Pareto Distribution (EGPD) introduced by Naveau et al. (2016). This distribution provides flexibility in both the bulk and the tail of the distribution, while having a tail behaviour similar to the Generalized Pareto Distribution (GPD) suggested by Extreme Value Theory. This family of distributions allows one to avoid explicit threshold selection and it is particularly efficient for precipitation modeling where we often need the full marginal distribution to model rainfall space-time aggregates leading to flood events. This distribution is a transformation of the classical GPD which corresponds to the limit distribution of threshold exceedances as the threshold $u$ tends to infinity with shape parameter $\xi$ and scale parameter $\sigma_u$. In a non-asymptotic setting, for a fixed threshold $u$, large enough, the corresponding cumulative distribution function is $H_\xi$ and we have the approximation

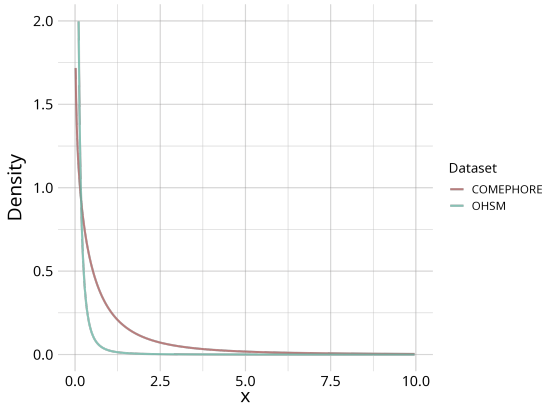
$$\mathbb{P}\left(X_s - u > y | X_s > u\right) \approx \overline{H}_\xi\left(\frac{y}{\sigma_u}\right) = \begin{cases} \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-1/\xi} & \text{if } \xi \neq 0\,, \\ e^{-\frac{y}{\sigma_u}} & \text{if } \xi = 0\,, \end{cases}$$

with $a_+ = \max(a, 0)$. The cumulative distribution function of the EGPD is given by adding a transformation $G$ to $H_\xi$. The efficient and simpler form is $G(x) = x^\kappa\,, \kappa > 0$. The parameter $\kappa$ parameter controls the lower tail behaviour. To reduce the impact of very low values, a small left censoring is added locally to better fit this distribution to real rainfall data. This censoring was chosen according to the smaller normalized root mean square error at each location (see Haruna et al. (2023)).

For both datasets described above, over similar time periods and similar space locations, we obtain a good fit of the EGPD margins. We have similar estimated parameters for all sites within the same dataset. The boxplots of the estimated EGPD parameters for all the locations and the theoretical EGPD density with mean parameter estimates are given in Figure 2. We have close tail parameter estimates for the two data sets but the hourly one gives a smaller $\xi$ estimate. Indeed, extreme events are smoothed within hourly data. Therefore, the tail of the distribution can appear heavier for OHSM data with a higher $\xi$ parameter. Then, we have a higher shape parameter estimate for the COMEPHORE data. This can be explained by the fact that these rain events are spread over a longer period of time and require a larger scaling parameter to fit the distribution. Finally, there is an expected difference between the $\kappa$ estimates, which is directly related to the time scale difference. In fact, a resolution of 5 minutes gives more low rainfall values for OHSM data, leading to a lower $\kappa$ parameter.




(a) Parameter estimates




(b) Mean density estimation

Figure 2: EGPD fit for both datasets over a similar period of 4 years. The COMEPHORE data correspond to 21 pixels around the location of the OHSM rain gauges. (a) boxplots of the parameter estimates for each dataset; (b) theoretical EGPD density with the mean of parameter estimates for each dataset.

# 3 Spatio-temporal dependence

## 3.1 Spatio-temporal process for rainfall modeling

Now that we have modeled the marginal behaviour of precipitation in the univariate analysis, the next step is to consider the spatio-temporal dependence. To understand the dependence structure of rainfall events, several standard asymptotic models exist such as max-stable and Pareto processes developed in Davison et al. (2012), Ferreira and De Haan (2014) and Dombry and Ribatet (2015). It seems that the Brown-Resnick process, introduced in Kabluchko et al. (2009), is the most commonly used for spatial and spatio-temporal modeling, especially for precipitation, with its dependence structure flexibility defined in terms of a variogram function. It is a stationary max-stable class of random fields based on Gaussian processes with stronger regularity properties that make it more suitable for modeling extreme events. Thanks to its structure, there are nice links with classical geostatistical properties.

Let $X = \{X_{\boldsymbol{s},t} \mid (\boldsymbol{s},t) \in \mathcal{S} \times \mathcal{T}\}$ be a strictly stationary isotropic Brown-Resnick process representing our spatio-temporal amounts of rainfall. Then, as defined in Buhl and Klüppelberg (2018), we have the following representation

$$X_{\boldsymbol{s},t} = \bigvee_{j=1}^{\infty} \xi_j e^{W^j_{\boldsymbol{s},t} - \gamma(\boldsymbol{s},t)}$$

where $\xi_j$ is a point of a Poisson process with intensity $\xi^{-2}d\xi$. The processes $W^j$ are independent replicates of $W = \{W_{\boldsymbol{s},t} \mid (\boldsymbol{s},t) \in \mathcal{S} \times \mathcal{T}\}$, an underlying Gaussian spatio-temporal process with stationary increments and with a semivariogram $\gamma$ defined by

$$2\gamma(\boldsymbol{s},t) = Var\left(W_{\boldsymbol{s},t} - W_{\boldsymbol{0},0}\right).$$

The semivariogram $\gamma$ can be used to determine the one of $X$. In fact, the Brown-Resnick process $X$ is defined by the underlying Gaussian process $W$, so that the spatio-temporal dependence structure of $X$ is directly determined by the one of $W$. So our goal is to find a way to model this measure of dependence in order to get information about the overall distributional dependence of $X$. In order to do this, we look at the space-time extremogram, which will be expressed as a function of the variogram. This measure gives information about the extreme spatio-temporal dependence of a spatio-temporal process by looking at the simultaneous excesses for each pair of locations separated by a spatial lag at two different times separated by a temporal lag. So, let $\Lambda_{\mathcal{S}} \subset \mathbb{R}_+^2$ and $\Lambda_{\mathcal{T}} \subset \mathbb{R}_+$ be sets of spatial and temporal lags respectively. Then, with the definition given in Coles et al. (1999), we can define the extremal dependence measure, in a spatio-temporal context, for $\boldsymbol{h} \in \Lambda_{\mathcal{S}}, \tau \in \Lambda_{\mathcal{T}}$ and $q \in [0, 1[$ as

$$\chi\left(\boldsymbol{h}, \tau\right) = \lim_{q \to 1} \chi_q\left(\boldsymbol{h}, \tau\right) , \quad \text{with} \quad \chi_q\left(\boldsymbol{h}, \tau\right) = \mathbb{P}(X^*_{\boldsymbol{s},t} > q \mid X^*_{\boldsymbol{s}+\boldsymbol{h},t+\tau} > q) ,$$

where for any $\boldsymbol{s} \in \mathcal{S}$ and any $t \in \mathcal{T}$, $X^*_{\boldsymbol{s},t}$ is the uniform rank transformation of $X_{\boldsymbol{s},t}$. As explained in Davis et al. (2013), the bivariate distribution of the max-stable process $X$ is given by

$$\mathbb{P}\left(X_{\boldsymbol{s},t} \leq x_1,\, X_{\boldsymbol{s}+\boldsymbol{h},t+\tau} \leq x_2\right) = \exp\left(-V_{\gamma(\boldsymbol{h},\tau)}\left(x_1, x_2\right)\right) , x_1 , x_2 > 0 ,$$

with

$$V_{\gamma(\boldsymbol{h},\tau)}\left(x_1, x_2\right) = \frac{1}{x_1}\phi\left(\frac{\log\frac{x_2}{x_1}}{2\sqrt{\frac{1}{2}\gamma\left(\boldsymbol{h},\tau\right)}} + \sqrt{\frac{1}{2}\gamma\left(\boldsymbol{h},\tau\right)}\right) + \frac{1}{x_2}\phi\left(\frac{\log\frac{x_1}{x_2}}{2\sqrt{\frac{1}{2}\gamma\left(\boldsymbol{h},\tau\right)}} + \sqrt{\frac{1}{2}\gamma\left(\boldsymbol{h},\tau\right)}\right)$$

where $\phi$ is the standard normal distribution function. For max-stable processes, this last function $V_\gamma(\boldsymbol{h}\,, \tau)$ has been linked to the extremogram. From Coles et al. (1999), we have

$$\chi\left(\boldsymbol{h}, \tau\right) \;\; = \;\; 2 - V_{\gamma(\boldsymbol{h},\tau)}\left(1, 1\right) .$$

Therefore, we obtain the spatio-temporal extremogram of the max-stable process of Brown-Resnick $X$, given by

$$\chi(\boldsymbol{h}, \tau) = 2\left(1 - \phi\left(\sqrt{\frac{1}{2}\gamma(\boldsymbol{h}, \tau)}\right)\right), \boldsymbol{h} \in \Lambda_{\mathcal{S}},\, \tau \in \Lambda_{\mathcal{T}}. \tag{1}$$

where $\phi$ is the standard normal distribution function and $\gamma$ is a stationary and isotropic variogram. With this expression we can obtain information about the overall distribution dependence over the spatio-temporal semivariogram $\gamma$.

## 3.2 Extremal dependence with additively separable variogram

We first model the spatio-temporal dependence with an additively separable structure, which facilitates parameter inference while providing a relatively simple but useful model that combines spatial and temporal dependence. To do this, we use the strategy developed in Buhl and Klüppelberg (2018).

We assume that we have a variogram with additive separability, so that we can get a linear parameterisation of the theorical extremogram with an appropriate transformation. Thus, we can write our spatio-temporal variogram, with $0 < \alpha_1, \alpha_2 \leq 2$ and $\beta_1, \beta_2 > 0$, as $\frac{1}{2}\gamma(\boldsymbol{h}, \tau) = \beta_1 \|\boldsymbol{h}\|^{\alpha_1} + \beta_2 \tau^{\alpha_2}$. Then, the theorical spatio-temporal extremogram of $X$ is given by

$$\chi\left(\boldsymbol{h}, \tau\right) = 2\left(1 - \phi\left(\sqrt{\beta_1 \|\boldsymbol{h}\|^{\alpha_1} + \beta_2 \tau^{\alpha_2}}\right)\right).$$

From equation (1), we have

$$\frac{1}{2}\gamma(\boldsymbol{h}, \tau) = \left(\phi^{-1}\left(1 - \frac{1}{2}\chi\left(\boldsymbol{h}, \tau\right)\right)\right)^2$$

and this leads to the transformation $\eta\left(\chi\right) = 2\log\left(\phi^{-1}\left(1 - \frac{1}{2}\chi\right)\right)$ which gives the following expressions with the variogram parameters. For the spatial case, by setting the time lag to 0, we obtain

$$\eta\left(\chi\left(\boldsymbol{h}, 0\right)\right) = \log\left(\beta_1\right) + \alpha_1 \log\left(\|\boldsymbol{h}\|\right) =: c_1 + \alpha_1 x_{\boldsymbol{h}}.$$

For the temporal case, by setting the spatial lag to $\mathbf{0}$, we obtain

$$\eta\left(\chi\left(\mathbf{0}, \tau\right)\right) = \log\left(\beta_2\right) + \alpha_2 \log\left(\tau\right) =: c_2 + \alpha_2 x_{\tau}.$$

Therefore, with the separability assumption, we can separate the spatial and the temporal cases into two linear expressions so that we can use a Weighted Least Squares Estimation (WLSE) for parameter inference on the spatial and the temporal empirical extremograms. A summary of the procedure is given by Figure 3.

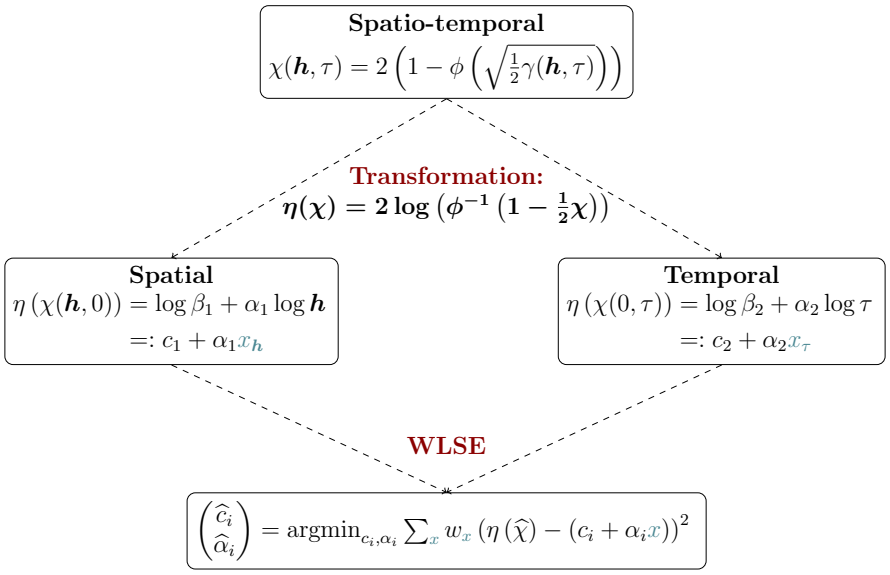



Figure 3: Procedure with the separability assumption on the spatio-temporal variogram

**Empirical spatial extremogram**

To estimate the spatial extremogram, we first define a spatial structure in an isotropic framework. For the smaller dataset, the OHSM one, we define a set of equifrequent classes $C_{\boldsymbol{h}}$ for each spatial lag $\boldsymbol{h} \in \Lambda_{\mathcal{S}}$ which leads us to the set of all pairs of locations within each distance class as

$$N(\boldsymbol{h}) = \left\{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in \mathcal{S}^2 \;\middle|\; \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \in C_{\boldsymbol{h}}\right\}.$$

For the largest COMEPHORE dataset, we have a spatial grid and we can define a set that contains all pairs of locations for the same spatial lag $\boldsymbol{h} \in \Lambda_{\mathcal{S}}$ as

$$N(\boldsymbol{h}) = \left\{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in \mathcal{S}^2 \;\middle|\; \|\boldsymbol{s}_i - \boldsymbol{s}_j\| = \|\boldsymbol{h}\|\right\}.$$

Then for a fixed $t \in \mathcal{T}$ and for any $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in N(\boldsymbol{h})$, a natural estimator of the spatial extremogram is

$$\widehat{\chi}_q^{(t)}(\boldsymbol{h}, 0) = \frac{\frac{1}{|N(\boldsymbol{h})|}\sum_{i,j\,|\,(\boldsymbol{s}_i,\boldsymbol{s}_j)\in N(\boldsymbol{h})} \mathbb{1}_{\left\{X^*_{\boldsymbol{s}_i,t}>q\,, X^*_{\boldsymbol{s}_j,t}>q\right\}}}{\frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} \mathbb{1}_{\left\{X^*_{\boldsymbol{s}_i,t}>q\right\}}}, \boldsymbol{h} \in \Lambda_{\mathcal{S}}$$

where $q$ is a high quantile (99.8%).

**Empirical temporal extremogram**

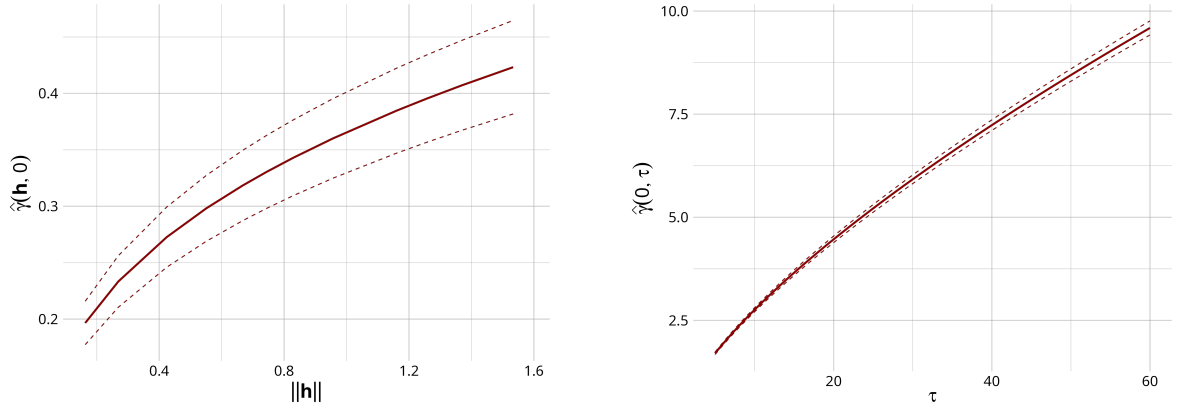
Regarding the temporal case, for a fixed $\boldsymbol{s} \in \mathcal{S}$ and for any $t \in \mathcal{T}$, a natural estimation of the temporal dependence measure is

$$\widehat{\chi}_q^{(\boldsymbol{s})}(0, \tau) = \frac{\frac{1}{T-\tau}\sum_{k=1}^{T-\tau} \mathbb{1}_{\{X^*_{\boldsymbol{s},t_k}>q\,, X^*_{\boldsymbol{s},t_k+\tau}>q\}}}{\frac{1}{T}\sum_{k=1}^{T} \mathbb{1}_{\{X^*_{\boldsymbol{s},t_k}>q\}}}, \tau \in \Lambda_{\mathcal{T}}$$

where $q$ is a high quantile (99.8%) and $t_k \in \{t_1, \ldots, t_T\} \subseteq \mathcal{T}$.

**Spatial and temporal variogram estimation**

We apply the described procedure on OHSM data to infer the variogram parameters. Then we obtain the final estimation of the spatial variogram $\widehat{\gamma}(\boldsymbol{h}, 0) = \widehat{\beta}_1\|\boldsymbol{h}\|^{\widehat{\alpha}_1}$ and the temporal variogram $\widehat{\gamma}(\boldsymbol{0}, \tau) = \widehat{\beta}_2\tau^{\widehat{\alpha}_2}$, illustrated in Figure 4. The spatial variogram has an exponential form, indicating a continuous decrease in spatial correlation with distance, without reaching an apparent range. This suggests, as expected, that the rainfall data continue to show some degree of spatial correlation for larger distances. However, even for very small distances, there is variability and a decrease in dependence. The temporal variogram shows an almost linear relationship, indicating temporal variability within the rainfall data. The linear form suggests that the dependence between observations decreases proportionally with the time even over short time periods. This was expected because rain clouds move with time and rain events can be brief. These variograms show the relevance of our model at such a fine spatio-temporal scale.




(a) Spatial variogram

(b) Temporal variogram

Figure 4: Estimation of the spatial and temporal variograms on OHSM data, where $\|\boldsymbol{h}\|$ units are kilometers and $\tau$ units are minutes

However, assuming a separable structure is a strong hypothesis in our modeling. We looked at the spatial variogram at different fixed time lags (not shown here) and see that a non-separable structure may be required. So we want to relax this hypothesis to get a better dependence model.

## 3.3 Adding advection to the dependence model

When dealing with precipitation, it may be relevant to consider the role of advection. In meteorology, this is the horizontal transport of properties such as heat or moisture by the movement of air masses such as wind and clouds. It can therefore influence the behaviour of precipitation. In fact, due to winds, there is a spatial and temporal movement of clouds and precipitation with a specific direction and speed. This can lead to characteristic dependence patterns in precipitation, which we can capture by appropriate parametric structures with the space-time variogram.

Let $\boldsymbol{V} \in \mathbb{R}^2$ be the advection vector. It can be defined by its polar coordinates $(S_V, \theta_V)$, where $S_V$ corresponds to a velocity and $\theta_V$ to a direction angle. As said in Lepioufle et al. (2012), if we consider the Lagrangian referential $L$, moving with the air masses, we have a link between the Lagrangian semivariogram and the Eulerian semivariogram with the advection: $\gamma_L(\boldsymbol{h}, \tau) = \gamma(\boldsymbol{h} - \boldsymbol{V}.\tau, \tau)$. Then if we add this quantity in our separable model presented in 3.2, we obtain

$$\frac{1}{2}\gamma(\boldsymbol{h} - \boldsymbol{V}.\tau, \tau) = \beta_1 \|\boldsymbol{h} - \boldsymbol{V}.\tau\|^{\alpha_1} + \beta_2 \tau^{\alpha_2}.$$

With this expression we have relaxed the previous strong hypothesis of separability and then we cannot use the WLSE approach to infer the 6 parameters of the spatio-temporal variogram. Instead, a parameter optimization method will be used.

Let $\boldsymbol{\Theta}$ be the vector of parameters to be estimated *i.e.* $\boldsymbol{\Theta} = (\beta_1, \beta_2, \alpha_1, \alpha_2, S_V, \theta_V)$ and let $m$ be the total number of space-time pair combinations. Let $E_p = E_{(\boldsymbol{s}_i, \boldsymbol{s}_j), (t_i, t_j)}$ be the indicator

of simultaneous excess for a pair of locations $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in \mathcal{S}^2$ and a pair of times $(t_i, t_j) \in \mathcal{T}^2$. We have $E_p = \mathbb{1}_{\{X^*_{\boldsymbol{s}_i,t_i} > q \,|\, X^*_{\boldsymbol{s}_j,t_j} > q\}}$ and then $\mathbb{E}(E_p) = \chi_{p,\boldsymbol{\Theta}}$. Hence $E_p$ follows a Bernoulli distribution with a probability parameter $\chi_{p,\boldsymbol{\Theta}}$. So we have $\sum_{p=1}^{m} E_p \sim \mathcal{B}(n_p\,;\, \chi_{p,\boldsymbol{\Theta}})$ where $n_p$ is the total number of records for the $p$-th combination. Hence, with $\underline{E} = (E_p)_{p\in\{1,\ldots,m\}}$, the log-likelihood in order to get parameter estimates is

$$\log\left(L_{\boldsymbol{\Theta}}\left(\underline{E}\right)\right) = \sum_{p=1}^{m}\left[\log\binom{n_p}{k_p} + k_p \log \chi_{p,\boldsymbol{\Theta}} + (n_p - k_p)\log(1-\chi_{p,\boldsymbol{\Theta}})\right],$$

The separability approach, described in 3.2, gives us an initial estimate of $\boldsymbol{\Theta}$ with a zero advection. These would be the initial parameters for the log-likelihood maximisation. We apply this procedure to both datasets to obtain an overall estimate of a constant advection over the study area and hence we obtain a spatio-temporal variogram estimate.

# 4 Perspectives

Until now, both rainfall data (OHSM and COMEPHORE) were studied separately. The first dataset corresponds to 17 random spatial points with a high spatio-temporal resolution, while in the other we have a regular spatial grid with a lower spatial and temporal resolution. In future work, we want to combine the two datasets to obtain more regular spatial point measurements across the study area, while maintaining the fine spatio-temporal resolution. To do this, we will explore a downscaling approach that will also allow us to compare the two datasets. Downscaling involves dividing each spatial and temporal intervals of the higher resolution dataset into a number of smaller intervals. This can be done by interpolating, or by using statistical or physical models.

In addition, we will look at other datasets, such as the basic climate data from Météo France[2], which provides rainfall amounts with a resolution of 6 minutes for a station in Montpellier. This dataset also provides wind information with magnitude and direction, which may be relevant for our analysis. Indeed, we currently have a dependence model with constant advection, which is a strong hypothesis. Taking this wind data into account in the model could lead to a more accurate estimation of the advection in space and time.

# References


Buhl, S., & Klüppelberg, C. (2018). Limit theory for the empirical extremogram of random fields. *Stochastic Processes and their Applications*, *128*(6), 2060–2082.

Coles, S., Heffernan, J., & Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, *2*, 339–365.

Davis, R. A., Klüppelberg, C., & Steinkohl, C. (2013). Max-stable processes for modeling extremes observed in space and time. *Journal of the Korean Statistical Society*, *42*(3), 399–414.


[2]The dataset is available here: https://meteo.data.gouv.fr/datasets/6569ad61106d1679c93cdf77

Davison, A. C., Padoan, S. A., & Ribatet, M. (2012). Statistical modeling of spatial extremes.

Dombry, C., & Ribatet, M. (2015). Functional regular variations, pareto processes and peaks over threshold. *Statistics and its Interface*, *8*(1), 9–17.

Ferreira, A., & De Haan, L. (2014). The generalized pareto process; with a view towards application and simulation.

Finaud-Guyot, P., Guinot, V., Marchand, P., Neppel, L., Salles, C., & Toulemonde, G. (2023). Rainfall data collected by the HSM urban observatory (OMSEV).

Haruna, A., Blanchet, J., & Favre, A.-C. (2023). *Modeling areal precipitation hazard: A data-driven approach to model intensity-duration-area-frequency relationships using the full range of non-zero precipitation in switzerland* (preprint). Preprints.

Kabluchko, Z., Schlather, M., & De Haan, L. (2009). Stationary max-stable fields associated to negative definite functions.

Lepioufle, J.-M., Leblois, E., & Creutin, J.-D. (2012). Variography of rainfall accumulation in presence of advection. *Journal of Hydrology*, *464-465*, 494–504.

Naveau, P., Huser, R., Ribereau, P., & Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*.

Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., & Laurantin, O. (2012). A 10-year (1997—2006) reanalysis of quantitative precipitation estimation over france: Methodology and first results. *IAHS-AISH publication*, (351), 255–260.

# Prédiction conforme

# On the Efficiency and Informativeness of Deep Conformal Classifiers using the Penalised Inverse Probability Nonconformity Function

Paul Melki [1,2], Lionel Bombrun [1,3], Boubacar Diallo [2],
Jérôme Dias [2] & Jean-Pierre Da Costa [1,3]

[1] *IMS, CNRS UMR 5218, Bordeaux INP, Université de Bordeaux, France*
[2] *EXXACT Robotics, France*
[3] *Bordeaux Sciences Agro, France*
**paul.melki@u-bordeaux.fr**

**Résumé.** Les prédictions conformes permettent de transformer n'importe quel predicteur ponctuel en un *prédicteur d'ensembles* avec des garanties formelles sur le recouvrement de la vraie classe à un niveau de confiance choisi. Un composant important de la chaîne de prédiction conforme est le *score de non-conformité* qui attribue à chaque observation une mesure "d'étrangeté" par rapport aux données vues précédemment. Plusieurs modèles conformes sont souvent comparés en fonction de leur *efficacité*, généralement mesurée par la taille moyenne des ensembles prédits, et leur *informativité*, le nombre de singletons prédits. Comme cela a été montré dans la littérature, ces deux critères sont influencés par les données, les performances du modèle de base et le score de non-conformité. Leur maximisation conjointe à l'aide d'une fonction de non-conformité ayant de bonnes propriétés est souhaitable. Le travail actuel introduit la fonction de non-conformité "Penalised Inverse Probability" (PIP) inspirée des fonctions de score classiques (*Hinge Loss* et *Margin Score*). À l'aide d'exemples illustratifs et de résultats empiriques en classification d'images de cultures en agriculture de précision, nous montrons que le PIP présente précisément le comportement souhaité, établissant un bon équilibre entre informativité et efficacité.

**Mots-clés.** Prédictions conformes, score de non-conformité, classification, incertitude.

**Abstract.** The conformal prediction framework transforms any point predictor into a *set predictor* with formal guarantees on the coverage of the true value at a chosen level of confidence. An important component of the conformal pipeline is the *nonconformity score function* which assigns to each observation a measure of "strangeness" in comparison to the previously seen data points. Multiple conformal models are often compared based on their *efficiency*, usually measured by the average size of the predicted sets, and their *informativeness*, the number of predicted singletons. As shown in the literature, these two criteria are influenced by the dataset, the performance of the base model and the nonconformity score function. The joint maximisation of these criteria using a well-behaved nonconformity function is desirable. The current work presents the "Penalised Inverse Probability" (PIP) nonconformity function inspired from classical score functions (Hinge Loss and Margin Score). Using some illustrative examples and experimental results on the task of crop and weed image classification in precision agriculture, we show that PIP exhibits precisely the desired behaviour, striking a good balance between informativeness and efficiency.

**Keywords.** Conformal prediction, nonconformity score, classification, uncertainty.

# 1 Introduction

Let $\mathbf{x} \in \mathcal{X}$ be a vector of features, which we will call an *object*, following the commonly used annotation in the conformal prediction literature (Vovk et al., 2005). For each object is associated a class label $y \in \mathcal{Y} := \{1, ..., K\}$ to form what we call an *example* $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. A black-box classifier $\mathcal{B}$ is trained on a set of $n_{\text{train}}$ examples to output for an individual a class prediction $\hat{\mathcal{B}}(\mathbf{x}) = \hat{y} \in \{1, ..., K\}$ and an associated estimated probability $\hat{p}^{\hat{y}} \in [0, 1]$, such that $\sum_{k=1}^{K} \hat{p}^k = 1$.

In inductive conformal prediction (Papadopoulos et al., 2002), the trained classifier is then calibrated on a held-out set of $n_{\text{cal}}$ calibration examples $\{(\mathbf{x}_i, y_i), i = 1, ..., n_{\text{cal}}\}$ using a well-defined nonconformity score function $\Delta(\hat{\mathcal{B}}(\mathbf{x})) = \Delta(\hat{y}) : \mathcal{Y} \to \mathbb{R}$. The nonconformity function assigns a "strangeness value" to each individual in the calibration set that measures how *conforming* it is to what the model has previously seen. The result of the calibration procedure is usually a quantile value $q_{\text{cal}}$ computed on the distribution of nonconformity scores over the calibration set. This quantile is used at the prediction phase to construct prediction sets $\mathcal{C}_{1-\alpha}(\mathbf{x}) \subset \mathcal{Y}$ that guarantee the inclusion of the true class $y$ at least $1 - \alpha$ of the times, for $\alpha \in (0, 1)$ a chosen significance level. This is known as the *marginal coverage* guarantee (Vovk et al., 2005):

$$\mathbb{P}\Big(y \in \mathcal{C}_{1-\alpha}(\mathbf{x})\Big) \geq 1 - \alpha \tag{1}$$

The strength of the conformal approach is that it does not assume any distribution of the data and is agnostic to the base model $\mathcal{B}$. Indeed, the coverage guarantee in Equation 1 requires only the weak assumption of exchangeability of the data (Aldous, 1985) to be valid over the distribution of all possible calibration sets (Stutz et al., 2022). However, while the coverage is guaranteed at $1 - \alpha$ regardless of $\mathcal{B}$, the nonconformity score function used and the dataset considered, the efficiency and informativeness of the conformal pipeline depend on these components.

Let us first define these two notions. Vovk et al. (2016) propose different criteria for measuring the efficiency of conformal predictors. In the current work, we consider the two most commonly used measures, studying the average size of predicted sets and the proportion of predicted singletons. For a test set $\{(\mathbf{x}_i, y_i), i = 1, ..., n_{\text{test}}\}$ of examples different than those taken for training the base classifier and conformal calibration, we can define:

- **Efficiency**, as the average size of the predicted sets:

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}_{1-\alpha}(\mathbf{x}_i)| \tag{2}$$

  which we would generally like to minimise without violating the coverage guarantee. Hence, for two conformal predictors guaranteeing coverage at the same $1-\alpha$ confidence level, the predictor producing smaller prediction sets is preferred, in general, and is considered more efficient.

- **Informativeness**, as the percentage of predicted sets of size 1 (often called *oneC* in the literature):

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\{|\mathcal{C}_{1-\alpha}(\mathbf{x}_i)|=1\}} \tag{3}$$

  where $\mathbb{1}_{\{\}}$ is the indicator function. Singletons are considered the most informative predictions and are the most useful in practice to construct decision rules. A most informative model would be one that predicts singletons only while guaranteeing marginal coverage.

The aim of this current work is to propose a novel nonconformity measure that optimises these two, often conflicting, criteria. Indeed, the experimental work by Johansson et al. (2017) studying the impact of different model-agnostic nonconformity functions – in particular, the Hinge Loss (Eq. 4) and the Margin Score (Eq. 5) – on single and ensembles of neural network classifiers has shown that neither of these score functions permits the joint maximisation of informativeness and efficiency. Their empirical results show that the Hinge Loss minimises the size of prediction sets, while the Margin Score maximises the number of singletons. In an attempt to reconcile these two nonconformity functions and optimise the two criteria, Aleksandrova and Chertov (2021) propose a conformal prediction algorithm that computes a prediction set using both nonconformity scores. A decision is then taken: if the Margin Score-based prediction set produces a singleton, it is taken as the final prediction. Otherwise, the prediction set obtained using the Hinge Loss is predicted, as it usually has the smaller size. The empirical evidence presented by the authors shows that their approach does not always provide better results than using any of the two classical nonconformity functions. Moreover, the algorithm requires repeating the calibration and the prediction steps twice using each of the score functions, which can be quite inefficient.

The current work introduces the "Penalised Inverse Probability" (PIP) nonconformity function computed using the estimated probabilities $\hat{p}^k$, $k \in \mathcal{Y}$. It can be interpreted, as will be shown, as both a regularised version of the Hinge Loss function and of the Margin Score. The aim of using this function is striking the balance between maximisation of informativeness and minimisation of inefficiency. Simple illustrative examples are presented to show the behaviour of this function under different possible configurations of output probabilities. Experimental results comparing the proposed measure with other functions from the literature, on the task of crop and weed image classification using deep neural networks for precision agriculture, show that PIP manifests precisely the type of balanced behaviour it is intended for.

# 2 Nonconformity Score Functions

## 2.1 Review of Some Nonconformity Scores

The nonconformity score function quantifies the "strangeness" of each new observation by implicitly comparing it to the "old" observations, previously seen during the training and

calibration of the predictive model (Shafer and Vovk, 2008). For the same base model $\mathcal{B}$, different nonconformity functions produce different conformal predictors. In this work, the following nonconformity score functions from the literature (Johansson et al., 2017; Romano et al., 2020; Angelopoulos et al., 2021) are compared. Let $y$ be a class of interest and $\hat{p}^y$ its estimated probability:

- **Hinge Loss (IP)** This score also known as *Inverse Probability* measures how far the estimated probability of $y$ is from the perfect score of 1:

$$\Delta^{\text{IP}}(y) = 1 - \hat{p}^y \tag{4}$$

  The score function measures the model's certainty in the class of interest. A minimum value is assigned to a class with maximum probability, while less probable classes are assigned higher values. The Hinge Loss is, in a sense, a very natural measure of nonconformity. However, it is important to notice that it does not take the scores assigned to other classes into consideration.

- **Margin Score (MS)** This score measures the difference between the estimated probability of $y$ and the highest estimated probability among the other classes:

$$\Delta^{\text{MS}}(y) = \max_{k \neq y} \hat{p}^k - \hat{p}^y \tag{5}$$

  In a sense, this score function measures the model's lack of confidence in class $y$. An implicit hypothesis assumed when using this score function is that good predictive models tend to assign the highest probability estimate to the true class. However, this is not always the case in many practical situations.

- **Regularised Adaptive Prediction Sets (RAPS)** This score function is first introduced by Romano et al. (2020) as part of their APS approach with the aim of constructing valid prediction sets that adapt to the difficulty of each observation. It was further extended by Angelopoulos et al. (2021) with the addition of a regularisation term to reduce the size of the predicted sets. This score function takes into consideration all the classes that have higher estimated probability than the class of interest. It is defined as the cumulative probability of $y$ plus a regularisation term.

  Let the operator $R(k)$ be the rank of class $k$ after the estimated probabilities $p^1, ..., p^K$ have been sorted in decreasing order, and $\hat{p}^{[r]}$ be the probability estimate of the class having rank $r$, such that $\hat{p}^k = \hat{p}^{[R(k)]}$, we can define the RAPS score function as:

$$\Delta^{\text{RAPS}}(y) = \underbrace{\sum_{r=1}^{R(y)-1} \hat{p}^{[r]} + u \cdot \hat{p}^{[R(y)]}}_{\text{APS}} + \underbrace{\lambda \cdot \Big(R(y) - k_{reg}\Big)^+}_{\text{regularisation}} \tag{6}$$

  Here, $u$ is uniform random variable on $(0, 1)$ for tie-breaking, while $\lambda$ (the penalisation amount) and $k_{reg}$ (the rank at which to start penalising) are regularisation parameters that can be fixed *a priori*, or optimised on a held-out dataset. Notice that for $\lambda = 0$ we have the non-regularised APS score. It is important to note that while RAPS aims for adaptivity and improved efficiency, it is not clear that it was intended to take the informativeness criterium into consideration.

## 2.2 Penalised Inverse Probability

This work introduces a new nonconformity function that combines ingredients of the three previously presented measures: *Penalised Inverse Probability* (PIP). Following the same notation presented previously, it is defined as:

$$\Delta^{\text{PIP}}(y) = \begin{cases} 1 - \hat{p}^y & \text{if } R(y) = 1 \\ \underbrace{1 - \hat{p}^y}_{\Delta^{\text{IP}}(y)} + \underbrace{\sum_{r=1}^{R(y)-1} \frac{\hat{p}^{[r]}}{r}}_{\text{regularisation}} & \text{otherwise} \end{cases} \tag{7}$$

The first part of the function is simply the well-known Hinge Loss function defined previously, which is blind, by default, to the estimated probabilities of the other classes. A regularisation term that consists of the cumulative probability of all the classes with a higher estimated probability than the class of interest, weighted by the inverse rank of the class, is added. This term alleviates one of the Hinge Loss' shortcomings, namely the fact that it does not take into consideration the estimated probabilities of other classes, without only considering the maximal class as in Margin Score. The penalisation term does assign the highest penalty to the maximal class, but also considers the classes in-between.

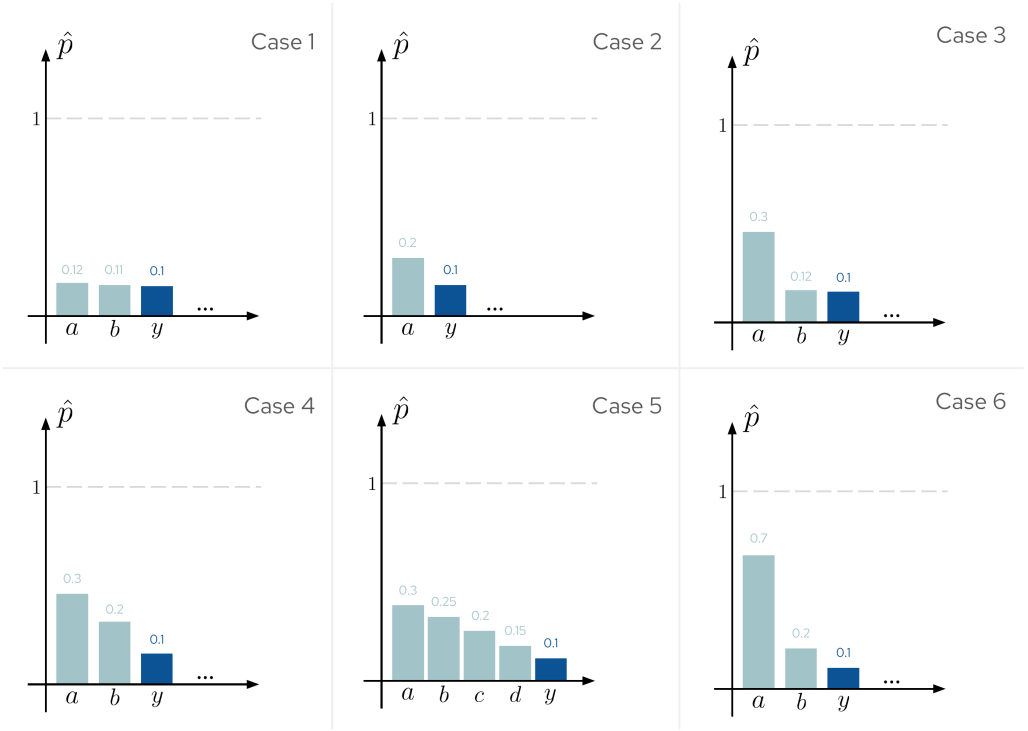



Figure 1: Six different potential configurations of model outputs sorted in decreasing order of $\hat{p}$. Only the classes until reaching the class of interest $y$ are shown. Computed nonconformity scores for each case can be seen in Table 1.

Consider the six different possible configurations shown in Figure 1. The class of interest being $y$, only the classes having estimated probabilities bigger than $\hat{p}^y = 0.1$ are shown, since they are the only ones taken into consideration for the computation of the different scores. Table 1 shows the different scores assigned by the IP, MS and PIP functions to the class $y$, sorted in increasing order (a higher score indicates a more uncertain class). The aim of this visualisation is to show how the proposed PIP function exhibits a more adaptive and balanced behaviour in different situations, unlike the classical, more rigid, IP and MS approaches that would assign the same score for quite distinct configurations.

Notice that the IP function, that takes into consideration only the estimated probability of the class of interest, assigns the same score to class $y$ in all configurations. The MS function assigns the smallest score to Case 1, since the difference between the maximum estimated probability and $\hat{p}^y$ is negligible, and attributes the highest score to Case 6 where the difference between these two values is quite large, although $y$ has the same rank in both cases. Case 2 is considered a bit "stranger" than Case 1 by the MS function because the difference between the maximal class and $y$ is a bit larger, which is an expected and desirable behaviour by this score function. However, Cases 3 to 5, albeit exhibiting quite different configurations, all have the same score. This is not surprising, but not desirable either: it would be important to distinguish between Case 3 where the difference between class $b$ and class $y$ is so negligible that class $y$ may very well have been predicted as the second class; and Case 4 where the difference between the second and third classes is more considerable, indicating that the model is even less confident about class $y$ in this case. Case 5, where the class of interest $y$ is quite far away from being in the top classes is, however, still assigned a similar MS value as Cases 3 and 4.

A more flexible behaviour is exhibited by the proposed PIP score function. A first glance at $\Delta^{\text{PIP}}(y)$ shows that it does not assign the same score to any of the six different cases, showing more versatility than the other two score functions. Case 1 is assigned the lowest score, because the difference in estimated probabilities between the three classes is considered negligible. Class $y$ in this case is not very "strange" because it could have been the first predicted class. Case 2 is slightly stranger because the difference away from the first class is larger and cannot be simply attributed to noise. Case 3 is assigned a higher score than Case 2 but a slightly smaller score than Case 4: $y$ is equidistant from the maximum class in both cases, but in Case 2 the difference between $b$ and $y$ is negligible and attributable to insignificant factors. Case 5 is further penalised because more significant classes have higher estimated probabilities than $y$. Case 6 is attributed the highest score because class $y$ can be considered highly "strange" in the current configuration, having a much lower $\hat{p}^y$ than the maximum class and being preceded by a class with significant difference. In brief, the PIP score function exhibits the following desirable behaviour:

- In all situations, the IP is a baseline value for the PIP function. As such, it will assign high scores for low probability classes, and lower scores for high probability classes. This kind of behaviour leads to lower average size of predicted sets since it tends to exclude the classes with low probability estimates.
- PIP allows to take into consideration the probability estimates of other classes, including the maximum probability class. In cases where $y$ has a low value compared to the

| | $\Delta^{\text{IP}}(y)$ | $\Delta^{\text{MS}}(y)$ | $\Delta^{\text{PIP}}(y)$ |
|---|---|---|---|
| Case 1 | 0.90 | 0.02 | **1.08** |
| Case 2 | 0.90 | 0.10 | **1.10** |
| Case 3 | 0.90 | 0.20 | **1.26** |
| Case 4 | 0.90 | 0.20 | **1.30** |
| Case 5 | 0.90 | 0.20 | **1.43** |
| Case 6 | 0.90 | 0.60 | **1.70** |

Table 1: Computed scores of the different example cases shown in Figure 1. The proposed $\Delta^{\text{PIP}}(y)$ manifests a more adaptive behaviour for the varying configurations than the classical IP and MS functions.

maximum class, the observation will be heavily penalised: a behaviour similar to that of MS. This behaviour may lead to more predicted singletons: in cases where one class is assigned a very high probability compared to others, all the other classes will be heavily penalised and thus excluded from the predicted sets.

- PIP also allows to distinguish the cases where the difference between the probability estimates of the class of interest and the other classes is significant or not, penalising less when such differences are negligible and can be attributed to some noise.

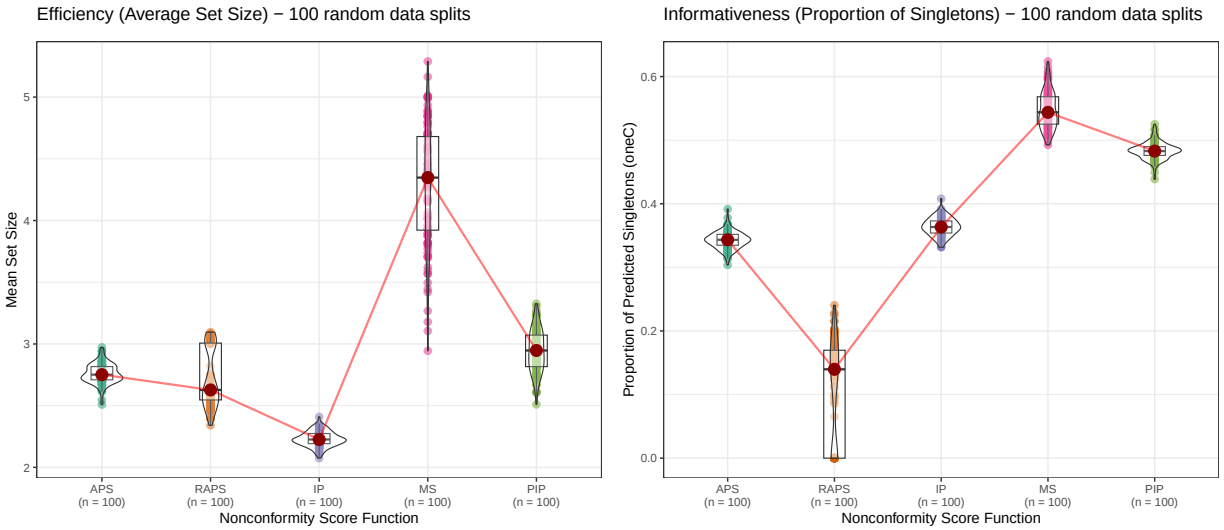
# 3 Comparison of Nonconformity Scores on Real Data

In this section, some empirical results studying the behaviour of the different nonconformity score functions for image classification are presented. These experiments are conducted in the context of a precision weeding application that aims at classifying images into different crop and weed species using deep neural networks (Melki et al., 2023).

## 3.1 Experimental Setup

The dataset considered consists of 14,800 RGB images distributed over 13 different classes obtained by dividing the original large images of the publicly available WE3DS dataset (Kitzler et al., 2023) into non-overlapping windows of size $224 \times 224$. Since the original large images also come with semantically annotated masks, the ground-truth class of each window is defined as the class with the highest number of pixels. The database is then randomly divided into: (1) a training set (70%), on which a ResNet18 classifier (He et al., 2016) is trained using default hyperparameters, and fixed for all experiments; the remaining 30% of the data are then split into (2) a calibration set (13.5%) for conformal calibration and (3) a test set (16.5%) on which the efficiency and informativeness of the conformal approach are evaluated.

After training the deep neural network, for each nonconformity function studied, the neural network is calibrated on the calibration set at the chosen confidence level of $1 - \alpha = 0.9$




(a) (b)

Figure 2: Boxplots of experimental results (each point is a random split): (a) *Efficiency*: PIP shows slightly larger average set sizes than APS, RAPS and Hinge – (b) *Informativeness*: PIP shows a significantly higher proportion of singletons compared to the other methods, and is competitive with the Margin Score.

and then used for constructing prediction sets on the test images. The calibration and prediction steps are repeated 100 times for each nonconformity function with a different random split of the data, in order to study the stability of the results. The RAPS hyperparemeters are fixed at $\lambda = 1$ and $k_{reg} = 3$ (Angelopoulos et al., 2021).

## 3.2 Results

As expected, all methods are able to maintain the required 90% marginal coverage guarantee on average. We compare the different nonconformity functions based on efficiency and informativeness. In Figure 2(a), we can see the average set size for each of the 100 runs for each nonconformity score function. The Hinge (IP) function leads to the smallest average set size and shows impressive stability over the different runs, which is in accordance with the results reported in (Johansson et al., 2017). RAPS manifests a smaller average set size than its non-regularised version (APS) (Angelopoulos et al., 2021). The Margin (MS) score function exhibits a very unstable behaviour over the different random runs, which manifests its deep dependence on the data considered. It also has a significantly higher average set size than all other methods, which echoes the results in (Johansson et al., 2017). Our proposed PIP score function is slightly less efficient than IP, APS and RAPS, but is still largely more efficient than MS.

A slight decrease in efficiency is a price to pay for the considerable gain in informativeness using PIP, as can be seen in Figure 2(b). Indeed, while MS manifests the highest proportion of singletons predicted, in accordance with Johansson et al. (2017), our proposed PIP

approach is not very away with around 50% of the predicted sets being singletons (all the while maintaining the coverage guarantee). The Hinge (IP) and APS show significantly lower informativeness, and RAPS shows the smallest number of singletons. Indeed, the proposed PIP score shows the intended behaviour of joint maximisation of the two criteria.

# 4 Conclusion

The current work introduced the "Penalised Inverse Probability" (PIP), a novel nonconformity function for conformal classification that can be used with any base model producing probability estimates for the predicted classes. The motivation behind PIP is the development of an elegant nonconformity function that jointly maximises efficiency and informativeness, while requiring minimal computational overhead. Using illustrative examples, the desirable behaviour of PIP in different conditions has been shown and compared to that of the classical Hinge Loss (IP) and Margin Score (MS) functions. Empirical experiments on the task of crop and weed image classification show promising results: PIP does indeed manifest the kind of behaviour it is intended for, that is, a good trade-off between maximising efficiency and maximising informativeness. During the oral presentation, further results comparing the behaviour of PIP and other nonconformity score functions on different datasets, multiple neural networks exhibiting varying levels of base accuracy, as well as other classification models will be exposed.

# References


Aldous, D. J. (1985). Exchangeability and related topics. In Hennequin, P. L., editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aleksandrova, M. and Chertov, O. (2021). Impact of Model-Agnostic Nonconformity Functions on Efficiency of Conformal Classifiers: An Extensive Study. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, pages 151–170. PMLR.

Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. (2021). Uncertainty Sets for Image Classifiers Using Conformal Prediction. In *International Conference on Learning Representations (ICLR)*, volume 2021.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Johansson, U., Linusson, H., Löfström, T., and Boström, H. (2017). Model-Agnostic Nonconformity Functions for Conformal Classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2072–2079.

Kitzler, F., Barta, N., Neugschwandtner, R. W., Gronauer, A., and Motsch, V. (2023). WE3DS: An RGB-D Image Dataset for Semantic Segmentation in Agriculture. *Sensors*, 23(5):2713.

Melki, P., Bombrun, L., Diallo, B., Dias, J., and Da Costa, J.-P. (2023). Group-Conditional Conformal Prediction via Quantile Regression Calibration for Crop and Weed Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 614–623.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive Confidence Machines for Regression. In Elomaa, T., Mannila, H., and Toivonen, H., editors, *Machine Learning: ECML 2002*, Lecture Notes in Computer Science, pages 345–356, Berlin, Heidelberg. Springer.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with Valid and Adaptive Coverage. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc.

Shafer, G. and Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12):371–421.

Stutz, D., Dvijotham, K. D., Cemgil, A. T., and Doucet, A. (2022). Learning Optimal Conformal Classifiers. In *International Conference on Learning Representations (ICLR)*, volume 2022.

Vovk, V., Fedorova, V., Nouretdinov, I., and Gammerman, A. (2016). Criteria of Efficiency for Conformal Prediction. In Gammerman, A., Luo, Z., Vega, J., and Vovk, V., editors, *Conformal and Probabilistic Prediction with Applications*, volume 9653, pages 23–39. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York, NY.

# Approximate full conformal prediction via influence functions in regression

Davidson Lova Razafindrakoto [1,2] & Alain Celisse [2] & Jérôme Lacaille [1]

[1] *Safran Aircraft Engines, France*
[2] *Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, France*
*Davidson-Lova.Razafindrakoto@edu.univ-paris1.fr*
*alain.celisse@univ-paris1.fr*
*jerome.lacaille@safrangroup.com*

**Résumé.** La prédiction conforme est un cadre qui fournit implicitement des régions prédictives de confiance sur la prédiction fournie par tout estimateur. Dans ce contexte, un problème majeur est le calcul de cette région prédictive de confiance, habituellement définie de façon implicite. Le point de vue "full conformal" où toute les observations d'entraînement sont utilisées pour cette région prédictive de confiance est d'ailleurs souvent abandonné pour cette raison au profit du point de vue "split conformal". La principale contribution de ce travail repose sur l'utilisation de fonctions d'influence afin de bâtir des régions prédictives de confiance approchées dans le cadre "full conformal" qui soient calculables efficacement. Parmi d'autres avantages, notre approche permet d'obtenir des régions prédictives de confiance plus informatives que celles obtenues par l'approche concurrente "split conformal". L'un des objectifs du présent travail consiste à bâtir des régions prédictives de confiance (approchées) pour la prédiction obtenue au terme de $T$ itérations d'un algorithme de type descente de gradient (GD). Les performances de notre approche seront illustrées au travers de modèles tels que : la régression linéaire, les réseaux de neurones de type MLP,...

**Mots-clés.** Apprentissage statistique, Réseaux de neurones, Régression, Quantification d'incertitudes, Prédiction conforme, Fonction d'influence

**Abstract.** Conformal prediction is the framework implicitly provides confidence predictive regions for the prediction of any estimator. In this context, a major issue is that the explicit computation of that predictive confidence region, is usually defined implicitly. The "full conformal" procedure where all training observations are used in building the confidence predictive region is often dismissed in favor of the "split conformal" procedure. The main contribution of this work lays on the use of influence function in order to build approximate confidence predictive regions in the "full conformal" framework which are efficiently calculable. Among other advantages, our approach enables to have, more informative confidence predictive regions than the competing approach "split conformal". One the of the aims of the present work consists at building (approximate) confidence predictive regions for the prediction given at the end of $T$ iterations of a gradient descent algorithm (GD). The performances of our approach will be illustrated through models such a: linear regression, neural networks (MLP),...

**Keywords.** Statistical learning, Neural networks, Regression, Conformal prediction, Influence function

# 1 État de l'art

La prédiction conforme est un cadre qui fournit implicitement des régions prédictives de confiance sur la prédiction fournie par tout prédicteur. Dans le cadre dit "full conformal", la région prédictive de confiance selon Vovk et al. (2005) est un sous-ensemble de l'ensemble des prédictions possibles en lesquelles la "p-valeur conforme" dépasse un seuil de confiance spécifié. Ces régions prédictives conformes dépendent de nombreux calculs de "p-valeurs conformes". Le nombre de calculs à effectuer est le cardinal de l'espace des prédictions possibles (Section 2.2.4 dans Vovk et al. (2022)).

Le calcul de cette "p-valeur conforme" requiert potentiellement plusieurs entraînements de prédicteurs, ce qui est coûteux sur le plan calculatoire. Cette difficulté est exacerbée par le nombre de calculs à effectuer. En particulier, c'est la raison pour laquelle l'approche "full conformal" est abandonnée au profit d'autres approches alternatives, notamment "split conformal" (Papadopoulos, 2008), "cross conformal", "Jackknife+" (Vovk, 2015),... Dans l'approche "split conformal", un seul entraînement est requis pour le calcul les "p-valeurs conformes" tandis que pour l'approche "cross conformal", le nombre d'entraînements peut aller d'un seul jusqu'au nombre total d'observations du jeu de données d'entraînement. Cependant, ces approximations motivées par les aspects calculatoires peuvent induire une perte d'informativité des régions fournies (Chapitre 4 de Vovk et al. (2022)).

Pour éviter cette perte, Nouretdinov et al. (2001) ont développé une méthode qui permet pour calculer efficacement les régions "full conformal" dans le cadre de la régression ridge. Récemment Ndiaye and Takeuchi (2019); Lei (2019) ont étendu ces méthodes pour d'autres variantes de la régression linéaire telles que Elastic Net et LASSO. Ndiaye (2022) utilise les propriétés de stabilité de l'entraînement pour approcher les régions "full conformal". Ndiaye and Takeuchi (2023) exploite la forme des régions prédictives de confiance et des algorithmes de recherche de racines pour les estimer. En classification, Martinez et al. (2023) utilisent les fonctions d'influence pour quantifier l'apport d'une nouvelle variable $y$ à prédire lorsqu'elle est incluse dans l'ensemble d'entraînement. Cette approximation est ensuite utilisée pour bâtir des régions prédictives de confiance approchées. Dans ce cadre, le prédicteur entraîné est celui fourni par la minimisation du risque empirique. Cependant plus généralement, le prédicteur entraîné est celui fourni au terme de $T$ itérations d'un algorithme d'optimisation de type descente de gradient (GD).

**Contributions.** Le présent travail traite de l'approximation par les fonctions d'influence de régions prédictives de confiance en régression dans le cadre "full conformal". Deux situations sont considérées ici : $(i)$ le prédicteur entraîné est celui qui minimise le risque empirique, $(ii)$ le prédicteur entraîné est celui fourni à l'issue de $T$ itérations de l'algorithme de descente de gradient. Ensuite pour chaque situation, les régions prédictives de confiance approchées sont explicitées dans le cas de la fonction de perte quadratique. Enfin, les approches développées sont évaluées au moyen de garanties théoriques de performances, de même qu'empiriquement par comparaison avec des approches concurrentes telles que l'oracle de Ndiaye and Takeuchi (2019), "split conformal" de Papadopoulos (2008) et "cross conformal" de Vovk (2015).

# 2 Prédiction conforme

## 2.1 Cadre “full conformal”

Soient $(X_1, Y_1), \ldots, (X_N, Y_N), (X_{N+1}, Y_{N+1})$ un ensemble d'observations *échangeables* (Section 2.1.5 de Vovk et al. (2022)). Pour un niveau de confiance $\alpha$, la région prédictive de confiance "full conformal" $\hat{C}_\alpha(X_{N+1})$ pour $Y_{N+1}$ a la garantie suivante (Vovk et al., 2005) :

$$\mathbb{P}(Y_{N+1} \in \hat{C}_\alpha(X_{N+1})) \geq 1 - \alpha.$$

Cette région $\hat{C}(X_{N+1})$ est définie à partir de l'entrée $X_{N+1}$ et des observations d'entraînement $(X_1, Y_1), \ldots, (X_N, Y_N)$ de la manière suivante :

$$\hat{C}(X_{N+1}) = \{y \in \mathcal{Y} : \pi_{N+1}(X_{N+1}, y) > \alpha\},$$

où $y \in Y$ est la prédiction à tester, et la p-valeur conforme $\pi_{N+1}(X_{N+1}, y)$ évaluée en le couple $(X_{N+1}, y)$ est donnée par

$$\pi_{N+1}(X_{N+1}, y) = \frac{1}{N+1}\text{Card}(\{i = 1, \ldots, N+1; S_i^y \geq S_{N+1}^y\}),$$

où $S_i^y$ et $S_{N+1}^y$ sont des scores de "non-conformité" définis ci-après (**Score**).

Posons $Z_i := (X_i, Y_i)$, $Z_{N+1}^y := (X_{N+1}, y)$ et

$$\mathcal{B}_{N+1}^{y;-i} = \rbag Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_N, Z_{N+1}^y \lbag \mathcal{B}_N = \rbag Z_1, \ldots, Z_N \lbag .$$

Remarquons que $\pi_{N+1}(X_{N+1}, y)$ est bâtie à partir d'une collection de $N+1$ scores dits de *non-conformité* $S_i^y$. Pour $i = 1, \ldots, N$, chaque score $S_i^y$ dépend de $(X_{N+1}, y)$ et de toute les autres observations sauf $(X_i, Y_i)$. Intuitivement, c'est une mesure de l'inadéquation entre la réponse cible $Y_i$ et la prédiction fournie par le prédicteur $M$ entraîné sur l'ensemble d'entraînement $\mathcal{B}_{N+1}^{y,-i}$

$$S_i^y = S(Y_i, M(X_i; \mathcal{B}_{N+1}^{y,-i})). \qquad \textbf{(Score)}$$

Le score $S_{N+1}^y$ quant à lui dépend de $(X_{N+1}, y)$ et de toutes les autres observations. C'est une mesure de l'inadéquation entre la prédiction à tester $y$ et la prédiction fournie par le prédicteur $M$ entraîné sur l'ensemble d'entraînement $\mathcal{B}_N$

$$S_{N+1}^y = S(y, M(X_{N+1}; \mathcal{B}_N))$$

avec $S$ une fonction appelée "mesure de non-conformité", par exemple le carré du résidu en régression (Kato et al. (2023); Balasubramanian et al. (2014), Section 2.9.5 de Vovk et al. (2022)).

En général, il n'existe pas de formule fermée qui permette de décrire le prédicteur $M(.; \mathcal{B}_{N+1}^{y,-i})$ entraîné sur $\mathcal{B}_{N+1}^{y,-i}$, en fonction de $y$ et de $i$. Il faut alors effectuer $N$ entraînements (autant que d'indices $i$) pour chaque prédiction $y$, ce qui est d'autant plus coûteux avec des prédicteurs "complexes". Le coût induit est ensuite multiplié par la taille de l'ensemble des prédictions $y$ possibles. Par exemple, pour $y \in \mathbb{R}$, le coût computationel est infini. C'est pour cela que l'approche "full conformal" est abandonnée au profit d'autres approches alternatives moins coûteuses telles que "split conformal" et "cross conformal" présentées dans les sections suivantes.

## 2.2 Cadre “split conformal”

La “p-valeur conforme” dans le cadre “split conformal” de Papadopoulos (2008) est calculée en deux temps. D'abord, une partie $\mathcal{B}_{\text{Train}}$ de l'ensemble d'entraînement est utilisée pour ajuster le prédicteur $M$. Ensuite, les scores de “non-conformité” sont évalués sur le sous-ensemble $\mathcal{B}_{\text{Calib}} = \mathcal{B}_N \setminus \mathcal{B}_{\text{Train}}$ et en $(X_{N+1}, y)$. Pour $(X_i, Y_i) \in \mathcal{B}_{\text{Calib}}$

$$S_i = S(Y_i, M(X_i; \mathcal{B}_{\text{Train}})) \text{ et } S_{N+1}^y = S(y, M(X_{N+1}; \mathcal{B}_{\text{Train}})).$$

La “p-valeur conforme” est donnée par

$$\pi_{N+1}^{\text{SCP}}(X_{N+1}, y) = \frac{\text{Card}(\{(X_i, Y_i) \in \mathcal{B}_{\text{Calib}} : S_i \geq S_{N+1}^y\}) + 1}{\text{Card}(\mathcal{B}_{\text{Calib}}) + 1}.$$

Un seul entraînement est fait sur $\text{Card}(\mathcal{B}_{\text{Train}})$ observations, et le nombre de scores de “non-conformité” est $\text{Card}(\mathcal{B}_{\text{Calib}})$. La région $\hat{C}_\alpha^{\text{SCP}}(X_{N+1})$ a alors la garantie

$$\mathbb{P}(Y_{N+1} \in \hat{C}_\alpha^{\text{SCP}}(X_{N+1})) \geq 1 - \alpha.$$

## 2.3 Cadre “cross conformal”

La “p-valeur conforme” dans la cadre “cross conformal” de Vovk (2015) est calculée en trois temps. D'abord, $\mathcal{B}_N$ est séparé en $K$ sous-ensembles $\mathcal{B}^k$ disjoints. Ensuite, pour chaque sous-ensemble $\mathcal{B}^k$, une “p-valeur” $\pi^k(X_{N+1}, y)$ est calculée comme dans “split conformal”

$$\pi_k(X_{N+1}, y) = \frac{\text{Card}(\{(X_i, Y_i)) \in \mathcal{B}^k : S_i^k \geq S_{N+1}^{y,k}\}) + 1}{\text{Card}(\mathcal{B}^k) + 1},$$

avec

$$S_i^k = S(Y_i, M(X_i; \mathcal{B}_N \setminus \mathcal{B}_k) \text{ et } S_{N+1}^{y,k} = S(y, M(X_{N+1}; \mathcal{B}_N \setminus \mathcal{B}_k).$$

Enfin, la “p-valeur conforme” $\pi_{N+1}^{K,\text{CV}}(X_{N+1}, y)$ est la moyenne des $\pi_k(X_{N+1}, y)$

$$\pi_{N+1}^{K;\text{CV}}(X_{N+1}, y) = \frac{1}{K} \sum_{k=1}^{K} \pi_k(X_{N+1}, y).$$

$K$ entraînements sont faits sur $N + 1 - \text{Card}(\mathcal{B}^k)$ observations pour un nombre de scores de “non-conformité” de $\text{Card}(\mathcal{B}^k)$. Pour $K = N$, la région $\hat{C}_\alpha^{\text{N;CV}}(X_{N+1})$ a la garantie suivante (Théorème 1 dans Barber et al. (2021))

$$\mathbb{P}(Y_{N+1} \in \hat{C}_\alpha^{\text{N;CV}}(X_{N+1})) \geq 1 - 2\alpha.$$

# 3 Approximation des scores par fonctions d'influence

Pour des prédicteurs "complexes", il n'existe en général pas de formule fermée pour $(y, i) \mapsto M(.; \mathcal{B}_{N+1}^{y,-i})$. En classification, Martinez et al. (2023) ont utilisé les fonctions d'influence pour construire une approximation du prédicteur $M(.; \mathcal{B}_{N+1}^{y,-i})$ à partir du prédicteur entraîné sur l'ensemble d'entraînement $M(.; \mathcal{B}_N)$ en quantifiant "l'influence" du rajout de $(X_{N+1}, y)$ et du retrait de $(X_i, Y_i)$ de $\mathcal{B}_N$ sur les scores de "non-conformité". Cette approximation a le mérite de *rendre accessible le calcul de la région prédictive de confiance approchée.*

## 3.1 Minimisation du risque empirique

Dans la suite, pour un modèle prédictif $M$ paramétré par $\theta$, le risque empirique $\hat{R}$ évalué sur $\mathcal{B}$ est défini par

$$\hat{R}_{\mathcal{B}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} l(y, M(x;\theta)) = \frac{1}{|\mathcal{B}|} \sum_{z\in\mathcal{B}} l(z;\theta),$$

avec $l$ une fonction de perte et $z = (x, y)$. Dans le cadre de la minimisation du risque empirique, le prédicteur $M(.; \mathcal{B})$ entraîné sur $\mathcal{B}$ est celui dont le paramètre $\hat{\theta}(\mathcal{B})$ minimise le risque empirique

$$\hat{\theta}(\mathcal{B}) \in \arg\min_{\theta\in\Theta} \hat{R}_{\mathcal{B}}(\theta).$$

Les scores de "non-conformité" pour le minimiseur du risque empirique sont donnés par

$$S_i^y = S(Y_i, M(X_i; \hat{\theta}(\mathcal{B}_{N+1}^{y,-i}))) = S(Z_i; \hat{\theta}(\mathcal{B}_{N+1}^{y,-i}))$$
$$S_{N+1}^y = S(y, M(X_{N+1}; \hat{\theta}(\mathcal{B}_N))) = S(Z_{N+1}^y; \hat{\theta}(\mathcal{B}_N)).$$

Martinez et al. (2023) définissent l'approximation $\tilde{S}_i^y$ du score $S_i^y$, à l'aide de la fonction d'influence d'une observation sur le score pour $i = 1, \ldots, N$ comme

$$S_i^y \approx \tilde{S}_i^y = S(Z_i; \hat{\theta}(\mathcal{B}_N)) + \mathcal{I}_{S;Z_i}(Z_{N+1}^y; \hat{\theta}(\mathcal{B}_N)) - \mathcal{I}_{S;Z_i}(Z_i; \hat{\theta}(\mathcal{B}_N)),$$

où l'influence $\mathcal{I}_{S;u}(z, \hat{\theta}(\mathcal{B}_N))$ du couple $z = (x, y)$ sur le score $S$ évalué en $u = (u_{\text{in}}, u_{\text{out}})$ est donnée par

$$\mathcal{I}_{S;u}(z; \hat{\theta}(\mathcal{B}_N)) = [\nabla_\theta s(u;\theta)]^T_{\theta=\hat{\theta}(\mathcal{B}_N)} I_{\hat{\theta};B_N}(z).$$

Ici, $I_{\hat{\theta};\mathcal{B}_N}(z)$ représente l'influence du couple $z = (x, y)$ sur le minimiseur du risque empirique $\hat{\theta}$ entraîné sur $\mathcal{B}_N$

$$I_{\hat{\theta};\mathcal{B}_N}(z) = -\frac{1}{N} \left[ (\nabla_\theta^2 \hat{R}_{\mathcal{B}_N}(\theta))^{-1} \nabla_\theta l(z;\theta) \right]_{\theta=\hat{\theta}(\mathcal{B}_N)}. \tag{1}$$

Remarquons que l'inverse de la hessienne du risque empirique intervient dans l'expression de $I_{\hat{\theta},\mathcal{B}_N}(z)$. Le calcul et l'inversion de cette matrice sont coûteux sur le plan computationel et d'autant plus à mesure que la dimension du paramètre $\theta$ augmente. Cependant, des approximations telles que K-FAC (Ba et al., 2016) et GGN (Gargiani et al., 2020) pourraient diminuer ce coût de sorte permettre d'obtenir des régions approchées pour des modèles hautement paramétrés du type réseaux de neurones profonds.

## 3.2 Descente de gradient

Pour des prédicteurs "complexes", en général nous ne disposons du minimiseur du risque empirique $\hat{\theta}(\mathcal{B})$. Le paramètre $\theta_T(\mathcal{B})$ n'est plus le minimiseur du risque empirique, mais est plutôt fourni au terme de $T$ itérations d'un algorithme d'optimisation de type descente de gradient appliquée à la minimisation de $\theta \mapsto \hat{R}_{\mathcal{B}}(\theta)$.

Les itérés successifs sont donnés par

$$\begin{aligned} & \theta_0 = \theta^{(0)} \\ \text{Pour } t = 0, \dots, T-1, \qquad & \theta_{t+1} = \theta_t - \eta_t [\nabla_\theta \hat{R}_{\mathcal{B}}(\theta)]_{\theta=\theta_t}. \end{aligned} \tag{2}$$

**Définition 1** *Dans ce cadre, les scores de "non-conformité" $S_i^y$ pour $i = 1, \dots, N+1$ sont donnés par*

$$\begin{aligned} S_i^y &= S(Y_i, M(X_i; \theta_T(\mathcal{B}_{N+1}^{y,-i}))) = S(Z_i; \theta_T(\mathcal{B}_{N+1}^{y,-i})) \\ S_{N+1}^y &= S(y, M(X_{N+1}; \theta_T(\mathcal{B}_N))) = S(Z_{N+1}^y; \theta_T(\mathcal{B}_N)). \end{aligned}$$

En appliquant la stratégie précédemment décrite basée sur les fonctions d'influence, l'approximation $\tilde{S}_i^y$ par fonctions d'influence du scores $S_i^y$ est donnée pour $i = 1, \dots, N$ par

$$S_i^y \approx \tilde{S}_i^y = S(Z_i, \theta_T(\mathcal{B}_N)) + \mathcal{I}_{S;Z_i}(Z_{N+1}^y; \theta_T(\mathcal{B}_N)) - \mathcal{I}_{S;Z_i}(Z_i; \theta_T(\mathcal{B}_N)).$$

où $\mathcal{I}_{S;u}(z; \theta_T(\mathcal{B}_N))$ de $z = (x, y)$ sur le score $S$ évalué en $u = (u_{\text{in}}, u_{\text{out}})$ est donnée par

$$\mathcal{I}_{S;u}(z; \theta_T(\mathcal{B}_N)) = [\nabla_\theta s(u, \theta)]_{\theta=\theta_T(\mathcal{B}_N)}^T I_{\theta_T;\mathcal{B}_N}(z).$$

Ici $I_{\theta_T;\mathcal{B}_N}(z)$ représente l'influence de l'observation $z = (x, y)$ sur l'itération $\theta_T$ de l'algorithme (2) qui lui, vise à minimiser le risque empirique $\hat{R}$ évalué sur $\mathcal{B}_N$. Il est calculer itérativement en parallèle avec la descente de gradient à l'aide de ses $T$ itérés $(\theta_t)_{t=0}^{T-1}$.

**Lemme 1** *L'influence $I_{\theta_T;\mathcal{B}_N}(z)$ de $z = (x, y)$ sur le paramètre $\theta_T$ entraîné sur $\mathcal{B}_N$ est donnée par*

$$I_{\theta_T;\mathcal{B}_N}(z) = \frac{1}{N} C_T(z),$$

*où $C_T(z)$ est l'itéré $T$ du schéma itératif suivant*

$$\begin{aligned} & C_1(z) = -\eta_0 [\nabla_\theta l(z; \theta)]_{\theta=\theta_0} \\ \textit{Pour } t = 2, \dots, T, \qquad & C_t(z) = (I - \eta_{t-1} [\nabla_\theta^2 \hat{R}_{\mathcal{B}_N}(\theta)]_{\theta=\theta_{t-1}}) C_{t-1}(z) \\ & \qquad - \eta_{t-1} [\nabla_\theta l(z; \theta)]_{\theta=\theta_{t-1}}. \end{aligned} \tag{3}$$

Remarquons que la matrice hessienne du risque empirique intervient à chaque itération. Cependant, contrairement à la méthode précédente (1), elle n'est pas inversée. La méthode peut alors être appliquée même quand la matrice hessienne est non-inversible.

*Éléments de preuve pour le lemme 1.* Posons $(\theta_{t+1}^{\epsilon,z})_{t=1}^{T}$ comme étant les itérés d'un algorithme d'optimisation de type descente de gradient appliquée à la minimisation de $\hat{R}_{\mathcal{B}_N}(\theta) + \epsilon l(z;\theta)$, le risque empirique $\hat{R}_{\mathcal{B}_N}(\theta)$ auquel est ajouté $\epsilon l(z,\theta)$. Les itérés successifs sont donnés par

$$\begin{aligned} \theta_0^{\epsilon,z} &= \theta^{(0)} \\ \text{Pour } t = 0,\dots,T-1, \qquad \theta_{t+1}^{\epsilon,z} &= \theta_t^{\epsilon,z} - \eta_t[\nabla_\theta(\hat{R}_{\mathcal{B}_N}(\theta) + \epsilon l(z;\theta))]_{\theta=\theta_t^{\epsilon,z}}. \end{aligned} \tag{4}$$

Pour $\epsilon = \frac{1}{N}$, $\theta_T^{\frac{1}{N},z}$ est le paramètre fourni à l'issue de $T$ itérations du schéma (2) le risque empirique à minimiser est évalué sur $\mathcal{B}_N \cup \lbag z \rbag = \lbag Z_1,\dots,Z_N,z \rbag$. $\theta_T^{\frac{1}{N},z}$ est approché par une approximation de Taylor d'ordre 1

$$\theta_T^{\frac{1}{N},z} = \theta_T^{0,z} + \frac{1}{N}\left[\frac{d}{d\epsilon}\theta_T^{\epsilon,z}\right]_{\epsilon=0} = \theta_T + \frac{1}{N}C_T(z).$$

Le schéma (3) est ensuite déduit de (4). Notons que le même raisonnement est appliqué pour $\mathcal{B}_N \setminus \lbag Z_i \rbag = \lbag Z_1,\dots,Z_{i-1},Z_{i+1},\dots,Z_N \rbag$ en prenant $z = Z_i$ et $\epsilon = -\frac{1}{N}$.

# 4 Régions prédictives de confiance approchées

## 4.1 Calcul des régions approchées

Nous supposons dans la suite que la fonction de perte $l$ et la fonction de score $S$ sont quadratiques. Dans ce cas, l'approximation des scores $\tilde{S}_i^y$ par fonctions d'influence est donnée pour $i = 1,\dots,N$ par

$$\tilde{S}_i^y = a_i y + b_i \text{ et } \tilde{S}_{N+1}^y = S_{N+1}^y = (y - M(X_{N+1};\theta))^2$$

où les coefficients $a_i$ et $b_i$ ont des expressions complètement connues qui dépendent des estimateurs envisagés (minimiseur du risque empirique ou descente de gradient). Nous obtenons la forme suivante pour l'approximation de la région conforme

$$\tilde{C}_\alpha(X_{N+1}) = \left(\bigcup_{k=0,\dots,K:p_k>\alpha} ]y_k, y_{k+1}[\right) \cup \left(\bigcup_{k=1,\dots,K:p_k>\alpha} \{y_k\}\right),$$

où la "p-valeur conforme" $p_k$ associée à l'intervalle $]y_k, y_{k+1}[$ est donnée pour $k = 0,\dots,K$ par

$$p_k = \frac{\text{Card}(\{i = 1,\dots,N : ]y_k, y_{k+1}[ \subseteq C_i\}|) + 1}{N+1},$$

et celle associée au singleton $\{y_k\}$ est donnée pour $k = 1,\dots,K$ par

$$p_k = \frac{\text{Card}(\{i = 1,\dots,N : y_k \in C_i\}|) + 1}{N+1}.$$

et enfin les $y_k$ sont définis à partir des $C_i$.

L'intervalle $C_i$ est celui sur lequel $\tilde{S}_i^y$ est supérieur à $S_{N+1}^y$. De manière équivalente celui sur lequel $P_i(y) := S_{N+1}^y - \tilde{S}_i^y$, défini dans le lemme 4.1, est négatif.

**Lemme 2** *Quand l'estimateur envisagé est le minimiseur du risque empirique, $P_i$ est un polynôme en $y$ qui est donné pour $i = 1, \ldots, N$ par*

$$P_i(y) = S_{N+1}^y - \tilde{S}_i^y = y^2 - (2M(X_{N+1}; \hat{\theta}(\mathcal{B}_N)) + a_i)y + M(X_{N+1}; \hat{\theta}(\mathcal{B}_N))^2 - b_i$$

*Quand l'estimateur envisagé est fourni à l'issue de $T$ itérations d'un algorithme d'optimisation de type descente de gradient, $P_i$ est un polynôme en $y$ qui est donné pour $i = 1, \ldots, N$ par*

$$P_i(y) = S_{N+1}^y - \tilde{S}_i^y = y^2 - (2M(X_{N+1}; \theta_T) + a_i)y + M(X_{N+1}; \theta_T)^2 - b_i$$

Ce qui implique que l'intervalle $C_i$ pour $i = 1, \ldots, N$ est donnée par

$$C_i = \begin{cases} [u_i, v_i] & \text{si } P_i \text{ admet deux racines réelles distinctes,} \\ \{w_i\} & \text{si } P_i \text{ n'admet qu'une seule racine réelle,} \\ \emptyset & \text{si } P_i \text{ n'admet pas de racines réelles.} \end{cases}$$

Enfin, $y_0 = -\infty$, $y_{K+1} = +\infty$ et $y_1, \ldots, y_K$ sont les $u_i, v_i, w_i$ ordonnés dans l'ordre croissant.

## 4.2 Illustration

Les $(X_i, Y_i)$ sont tirées indépendamment et sont données par

$$X_i \sim \mathcal{U}([-1, 1]) \text{ et } Y_i = f(1.8(X_i + 0.8)) + \epsilon_i,$$

où $\epsilon_i \sim \mathcal{N}(0, 0.1)$ et

$$f : x \mapsto f(x) = 3\exp\left(-\frac{(x-1.8)^2}{0.28}\right) + 1.6\exp\left(-\frac{(x-0.7)^2}{0.13}\right) + 2\exp\left(-2.8x\right).$$

Le prédicteur est un réseau de neurones à une couche cachée à 10 neurones, ajusté par 200 itérations de la descente de gradient avec un pas de 0.05. "Split conformal" (a) est appliqué avec 100 observations d'entraînement, 100 de calibration. Les approches 3.1 (b) et 3.2 (c) sont appliqués avec 200 observations d'entraînement. Pour chaque méthode, les régions à niveau de confiance $1 - \alpha = 90\%$ sont évaluées sur 200 points de validation.

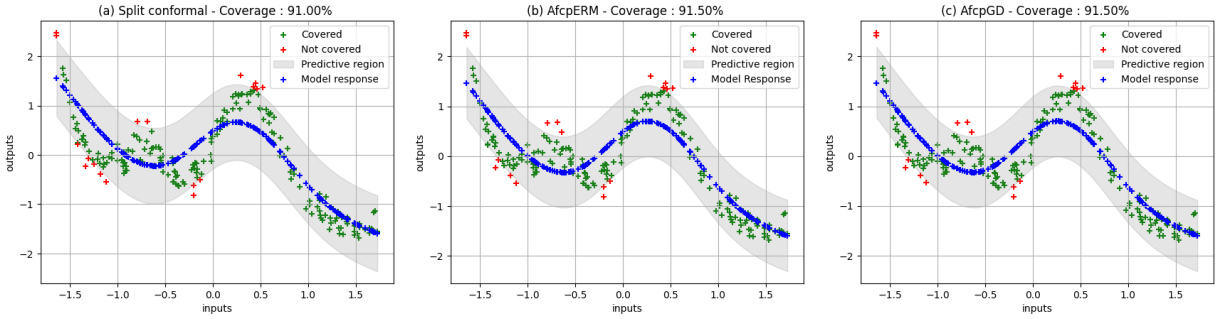



Figure 1: Régions prédictives et taux de couverture

Remarquons que pour chaque méthode, la figure 1 montre que le taux de couverture, c'est-à-dire la proportion de points de validation inclus dans les régions prédictives (ici en vert) est proche de $1 - \alpha$, le niveau de confiance spécifié.

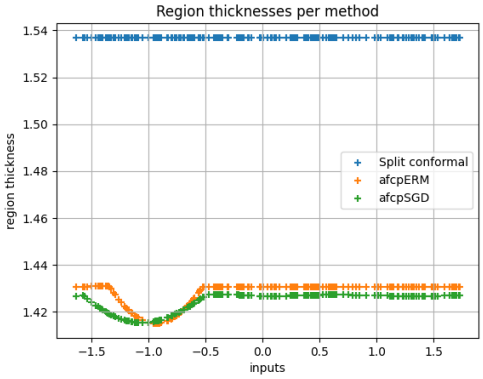


Figure 2: Épaisseur des régions par méthode

De plus, la figure 2 montre que la méthode 3.2 donne des régions moins épaisses que celles de la méthode 3.1, qui elles sont moins épaisses que celles données par "split conformal". Cependant, il faudra davantage d'expérimentations pour trancher.

# 5 Discussion

Le présent travail sera suivi d'une étude théorique et empirique de la précision l'approximation des scores, la précision de l'approximation des p-valeurs, le taux de couverture (validité) et la taille des régions prescrites (précision) en fonction de la taille de l'échantillon d'entraînement. Ces performances seront illustrées suivant des critères d'intérêt tels que la "complexité" du modèle prédictif, la dimension des covariables, la valeur du paramètre de régularisation dans le risque empirique et la valeur du seuil de confiance. L'approche proposé sera enfin comparée au prédicteur conforme oracle de Ndiaye (2022), "split conformal" de Papadopoulos (2008) et "cross conformal" de Vovk (2015). Nous envisageons aussi d'effectuer ces évaluations quand la matrice hessienne est remplacée par une approximation (Gargiani et al., 2020; Ba et al., 2016).

# Bibliographie

Ba, J., Grosse, R., and Martens, J. (2016). Distributed second-order optimization using kronecker-factored approximations. In *International Conference on Learning Representations*.

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+.

Gargiani, M., Zanelli, A., Diehl, M., and Hutter, F. (2020). On the promise of the stochastic generalized gauss-newton method for training dnns. *arXiv preprint arXiv:2006.02409*.

Kato, Y., Tax, D. M., and Loog, M. (2023). A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383.

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4):749–764.

Martinez, J. A., Bhatt, U., Weller, A., and Cherubin, G. (2023). Approximating full conformal prediction at scale via influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6631–6639.

Ndiaye, E. (2022). Stable conformal prediction sets. In *International Conference on Machine Learning*, pages 16462–16479. PMLR.

Ndiaye, E. and Takeuchi, I. (2019). Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32.

Ndiaye, E. and Takeuchi, I. (2023). Root-finding approaches for computing conformal prediction set. *Machine Learning*, 112(1):151–176.

Nouretdinov, I., Melluish, T., and Vovk, V. (2001). Ridge regression confidence machine. In *ICML*, pages 385–392. Citeseer.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Second edition.

# Prévision Conforme adaptative avec un pas de gradient explicite

Guillaume Principato[1,2,*]

[1] *Laboratoire de Mathématiques d'Orsay (LMO), Université Paris-Saclay, France*
[2] *EDF R&D, France*
[*] *guillaume.principato@universite-paris-saclay.fr*

**Résumé.** La prévision conforme est une méthode qui permet de construire des intervalles de prévision théoriquement valides à partir d'une prévision boîte noire. Ces dernières années, des méthodes ont été proposées pour l'appliquer à des séries temporelles qui, par définition, ne vérifient pas l'hypothèse d'échangeabilité nécessaire à la théorie initiale. Nous présentons les méthodes adaptatives de la littérature en insistant sur l'interprétation des garanties que l'on peut obtenir. Ensuite, nous montrons de nouveaux résultats sur un algorithme de descente de gradient à pas de gradient adaptatif qui n'avait, jusqu'alors, qu'été utilisé comme intermédiaire. Nous verrons, en le comparant aux autres algorithmes adaptatifs, que son utilisation en tant que tel est justifiée.

**Mots-clés.** Prévision conforme, inférence conforme adaptative, série temporelle, optimisation convexe en ligne.

**Abstract.** Conformal prediction is a method for constructing theoretically valid prediction sets from any black-box forecaster. In recent years, methods have been proposed to apply it to time series which, by definition, do not verify the exchangeability assumption required by the initial theory. We present adaptive methods from the literature, focusing on the interpretation of the guarantees that can be obtained. Next, we show new results on an gradient descent algorithm with adaptive step size that had hitherto only been used as an intermediary. By comparing it with other adaptive algorithms, we show that its use as such is justified.

**Keywords.** Conformal prediction, adaptive conformal inference, time series, online convex optimization.

## 1 Introduction

Considérons une série temporelle $(X_t, Y_t)_{t\in[\![1,T]\!]}$ avec $Y_t \in \mathbb{R}$ la variable que l'on cherche à prédire et $X_t \in \mathbb{R}^m$ un ensemble de $m$ covariables qui expliquent $Y_t$. L'objectif est de construire un intervalle de prévision qui a une probabilité $1-\alpha$ de contenir l'observation $Y_t$, où $\alpha$ est un paramètre que l'utilisateur choisit en fonction du type d'intervalle qu'il souhaite construire. Le cadre général des prévisions conformes introduit par Vovk et al. (2005) permet cela dans le cadre échangeable.

Pour ce faire, il faut construire une distribution empirique de scores, que l'on note $\mathcal{D}_t$, pour ensuite obtenir un intervalle en sélectionnant $y$ tel que son score soit plus petit qu'une proportion $1-\alpha$ des scores : $\hat{C}_t(\alpha) := \{y : S_t^y \leq \text{ Quantile}\,(1-\alpha, \mathcal{D}_t)\}$. Par exemple, si on note $\hat{\mu}(X_t)$ la prévision de la moyenne associée à $Y_t$ à partir des covariables $X_t$ on peut choisir comme score $S_t^y = |\hat{\mu}(X_t) - y|$.

L'inférence conforme adaptative (Adaptive Conformal Inference, ACI) est une méthode proposée par Gibbs and Candes (2021) qui se fonde sur le schéma classique de la prévision conforme mais qui intègre un aspect adaptatif. Cela permet alors d'obtenir des résultats même lorsque l'hypothèse d'échangeabilité n'est pas vérifiée.

Dans ce résumé, nous présentons différentes variantes d'algorithmes de type ACI introduites récemment par la littérature, en détaillant des critères qui attestent de la performance de ces algorithmes. La contribution principale est l'étude plus approfondie d'un algorithme à pas de gradient adaptatif (Algorithme 2). Nous montrons qu'il vérifie simultanément un ensemble de critères.

## 1.1 Présentation de l'algorithme ACI

Sans aucune hypothèse sur la distribution des observations $Y_t$, il est impossible d'avoir des garanties sur la probabilité de couverture de l'intervalle de prévision $\hat{C}_t^\alpha$ pour chacune des observations. Cependant, la littérature considère alors l'hypothèse suivante :

$$\text{Il existe un } \alpha_t^* \text{ tel que : } \mathbb{P}\left(Y_t \in \hat{C}_t(\alpha_t^*)\right) = 1-\alpha \tag{1}$$

où $\alpha_t^*$ et $\alpha$ ne sont pas nécessairement égaux ce qui permet de modéliser les changements de lois des $Y_t$. L'idée est alors de ne plus considérer l'intervalle $\hat{C}_t(\alpha)$ mais $\hat{C}_t(\alpha_t)$ avec $\alpha_t$ qui est appris sur les données afin de pouvoir s'adapter à un potentiel changement dans la distribution. Pour ce faire, Gibbs and Candes (2021) proposent une formule récurrente très interprétable :

**Algorithme 1.** ***(ACI)***

$$\alpha_{t+1} = \alpha_t + \gamma\,(\alpha - \text{err}_t) \qquad \textit{avec } \text{err}_t := \begin{cases} 0, & \textit{si } Y_t \in \hat{C}_t\,(\alpha_t) \\ 1, & \textit{si } Y_t \notin \hat{C}_t\,(\alpha_t) \end{cases} \tag{2}$$

Ainsi, si l'observation à l'instant précédent n'est pas couverte par l'intervalle, c'est probablement que ce dernier est trop étroit. Par conséquent, on l'agrandit en regardant un $\alpha_{t+1} < \alpha_t$. A l'inverse, si l'intervalle couvre l'observation, c'est qu'il est probablement trop large et donc $\alpha_{t+1} > \alpha_t$, ce qui réduit la largeur de l'intervalle. La formule récurrente (2) peut être vue comme un pas de descente de gradient par rapport à la fonction de perte quantile (3).

## 1.2 Garanties et limitations

L'hypothèse d'échangeabilité n'étant pas vérifiée dans le cas des séries temporelles, on passe d'un critère de validité en probabilité (Vovk et al., 2005) à un critère empirique de validité asymptotique.

**Critère 1.** ***(Validité)***
*Un intervalle de prévision* $\left(\hat{C}_t^{\alpha}\right)_{t\in[\![1,T]\!]}$ *est dit valide asymptotiquement si sa couverture asymptotique vaut* $1-\alpha$ *:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{Y_t\in\hat{C}_t^{\alpha}\}}\overset{p.s.}{=}1-\alpha$$

Gibbs and Candes (2021) montrent que l'intervalle produit par ACI est valide asymptotiquement avec une vitesse de convergence en $O\left(1/T\right)$ pour tout $\gamma\in\mathbb{R}^{\star}$ fixe, ce qui est la meilleure vitesse parmi les algorithmes du même type. Ce résultat étant valable pour tout $\gamma\in\mathbb{R}^{*}$ fixe, il ne donne pas de moyen de déterminer la valeur du pas de gradient $\gamma$. Ce paramètre est pourtant très important en pratique comme le soulignent Gibbs and Candes (2021). Une manière de déterminer $\gamma$ pourrait être de le contraindre par un second critère. En l'occurrence, notons $\beta_t := \sup\{\beta : Y_t \in \hat{C}_t(\beta)\}$ le plus grand niveau de quantile tel qu' $Y_t$ soit dans l'intervalle de prévision et la perte quantile de niveau $1-\alpha$ :

$$\ell(\beta_t,\theta) := (\alpha-\mathrm{err}_t)(\beta_t-\theta) \tag{3}$$

On remarque que $\nabla_\theta\ell\left(\beta_t,\alpha_t\right)=\mathrm{err}_t-\alpha$. Ainsi, (2) s'écrit également $\alpha_{t+1}=\alpha_t-\gamma\nabla_\theta\ell\left(\beta_t,\alpha_t\right)$, ce qui correspond à une étape de descente de gradient. Par conséquent, il est possible d'utiliser les méthodes classiques d'optimisation convexe en ligne pour obtenir des garanties théoriques comme, par exemple, des bornes de regret sur la perte quantile.

**Critère 2.** ***(Efficacité)***
*On définit le regret d'un algorithme de type ACI par rapport à la perte quantile par :*

$$\mathrm{Reg}_T(ACI)=\sum_{t=1}^{T}\ell(\beta_t,\alpha_t)-\inf_{\theta}\ell(\beta_t,\theta)$$

*On considère alors qu'un algorithme est performant lorsque ce regret peut être borné par un terme sous-linéaire.*

Les travaux de Zinkevich (2003) suggèrent alors de choisir $\gamma$ fixe tel que $\gamma=O\big(1/\sqrt{T}\big)$ pour (2) ce qui permet d'obtenir $\mathrm{Reg}_T(ACI)\leq O\big(\sqrt{T}\big)$. Cependant, cette approche n'est pas satisfaisante car cela ne garantit pas de bons résultats à tout instant, mais uniquement pour un horizon $T$ fixé.

# 2 Améliorations d'ACI

Nous avons conclu la partie précédente en insistant sur le fait que le choix du pas de gradient $\gamma$ était un enjeu majeur dans la construction d'un algorithme de type ACI. La littérature a proposé dans un premier temps de choisir une valeur constante, et a étudié dans un second temps des valeurs adaptatives (changeant au cours du temps, et déterminées selon les observations), ce dont nous rendons compte dans cette partie.

## 2.1 Méthodes d'agrégation d'experts

Une façon d'obtenir un pas de gradient $\gamma$ qui varie en fonction de l'instant est de procéder par agrégation d'experts. Il s'agit de définir une grille de $\gamma$ candidats et de considérer une instance d'ACI pour chaque $\gamma$. On obtient alors un ensemble d'experts $\{\alpha_t^\gamma\}$ que l'on agrège en considérant comme fonction de coût la perte quantile (3). Cette méthode est celle choisie par Zaffran et al. (2022) et Gibbs and Candès (2022) avec respectivement les algorithmes Online Expert Aggregation on ACI (AgACI) et Dynamically-tuned ACI (DtACI).

Ces algorithmes donnent d'excellents résultats pratiques (même si cette façon de construire l'intervalle rend difficile l'obtention d'une garantie par rapport au critère 1). Pour expliquer cette réussite, Gibbs and Candès (2022) mettent en avant le fait que leur algorithme produit des niveaux de quantile $1-\alpha_t$ qui sont proches des niveaux $1-\alpha_t^*$[1]. Plus formellement, on peut introduire le critère suivant :

**Critère 3.** ***(Efficacité)*** *Soit $\alpha_t^*$ tel que $\mathbb{P}\left(Y_t \in \hat{C}_t\left(\alpha_t^*\right) \mid \{\beta_s\}_{s<t}\right) = 1-\alpha$ et $\alpha_t$ produit par un algorithme de type ACI. On souhaite pouvoir contrôler le terme :*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\left[\left(\alpha_t-\alpha_t^*\right)^2\right]}{2}$$

*avec l'espérance qui porte sur $Y_t$ et les $(Y_s)_{s\in[\![1,t-1]\!]}$.*

Ce critère valorise les algorithmes qui s'adaptent rapidement aux changements de distribution et permet, sous réserve d'hypothèses supplémentaires, d'obtenir des garanties plus interprétables.

[1] Le niveau $\alpha_t^*$ défini dans le critère 3 n'est pas en adéquation avec sa précédente définition (1) pour laquelle il n'y avait pas de conditionnement par rapport aux $(\beta_s)_{s<t}$. Nous conservons cette double notation pour nous conformer au reste de la littérature.

## 2.2 Algorithmes à pas de gradient $\gamma_t$ explicite

Des approches différentes sont proposées par Bhatnagar et al. (2023) et Angelopoulos et al. (2024) pour définir séquentiellement le paramètre $\gamma_t$. Dans les deux cas, l'idée est de proposer un $\gamma_t$ qui, dans un premier temps, est suffisamment grand pour s'adapter à un potentiel changement de loi des observations $Y_t$, puis, dans un second temps, décroît lentement pour permettre la convergence. Bhatnagar et al. (2023) utilisent un algorithme qu'ils nomment Scale-Free Online Gradient Descent (SF-OGD) qui se caractérise par un pas $\gamma_t = O\big(1/\sqrt{t}\big)$ inspiré du pas $\gamma_t$ qui minimise le regret à tout instant, tandis que Angelopoulos et al. (2024) proposent une version plus générale avec $\gamma_t$ décroissant.

Ces derniers montrent qu'en prenant un tel $\gamma_t$, il est possible d'obtenir simultanément de bonnes garanties dans le cas adversarial ainsi que dans le cas *i.i.d.* tout en montrant que cela aurait été impossible à $\gamma$ fixe. Ces résultats récents sont dans le même esprit que la contribution de cet article qui met en valeur le fait que l'algorithme de descente de gradient à pas adaptatif vérifie simultanément différents critères.

# 3 Descente de gradient à pas adaptatif

Dans cette partie, nous présentons de nouveaux résultats obtenus sur un algorithme déjà proposé dans la littérature de la prévision conforme adaptative par Bhatnagar et al. (2023). Nous montrons qu'il satisfait à la fois un critère de validité (critère 1) et des critères d'efficacité (critères 2 et 3). Ceci constitue une nouveauté dans le sens où aucun des algorithmes proposés précédemment dans la littérature ne satisfait à la fois le critère 1 et 3.

**Remarque :** Bhatnagar et al. (2023) proposent cet algorithme dans l'optique de servir d'intermédiaire à un algorithme fortement adaptatif (Orabona, 2019) (Strongly Adaptive Online Conformal Prediction, SAOCP). Ces derniers n'ont donc pas étudié en détail l'algorithme de descente de gradient à pas adaptatif, ce que nous faisons ici.

## 3.1 Présentation de l'algorithme

D'abord appliqué aux prévisions conformes sous le nom SF-OGD, l'algorithme de descente de gradient à pas adaptatif introduit par Zinkevich (2003) pour l'optimisation convexe séquentielle, est une façon d'obtenir un pas de gradient $\gamma_t$ qui ait une expression explicite et donc pour laquelle il est plus simple d'obtenir de bonnes garanties théoriques.

**Algorithme 2.**

$$\alpha_{t+1} = \alpha_t + \frac{\eta}{\sqrt{\sum_{s=1}^{t} \nabla_\theta \ell\left(\beta_s, \alpha_s\right)^2}} \left(\alpha - \mathrm{err}_t\right) \tag{4}$$

*Notons $D$ le maximum des $\alpha_t$. On peut choisir $\eta = D/\sqrt{2}$ pour minimiser la borne du regret à tout instant (Orabona, 2019).*

**Remarque :** La formule (4) revient à écrire $\gamma_t = \dfrac{\eta}{\sqrt{\sum_{s=1}^{t} \nabla_\theta \ell\left(\beta_s, \alpha_s\right)^2}}$, ce qui est un $O\left(\frac{1}{\sqrt{t}}\right)$.

## 3.2 Garantie de validité

Comme souligné par Angelopoulos et al. (2024), sous réserve d'avoir des $\gamma_t$ décroissants, un algorithme construit d'après la formule (2) répond au critère 1. Nous vérifions cela en nous inspirant de la preuve de Gibbs and Candes (2021) pour l'algorithme 2 :

**Proposition 1.** *Prenons la convention $\hat{Q}_t(x) = -\infty$ pour $x < 0$ et $\hat{Q}_t(x) = +\infty$ pour $x > 1$. Alors, avec probabilité 1, on a pour tout $T \in \mathbb{N}$,*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t - \alpha \right| \leq O\left(\frac{1}{\sqrt{T}}\right)$$

*et donc*

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{Y_t \in \hat{C}_t^\alpha\}} \overset{p.s.}{=} 1 - \alpha$$

La contrepartie par rapport au cas à $\gamma$ fixe est que l'on perd en vitesse de convergence ($1/T$) (il est à noter que la vitesse obtenue par Bhatnagar et al. (2023) était encore pire, de l'ordre de $T^{-1/4}$ à facteurs logarithmiques près). Une question que nous voulons continuer d'explorer est l'amélioration de cette vitesse de convergence dans le cas adaptatif.

## 3.3 Garanties d'efficacité

L'algorithme 2 est en réalité défini de telle sorte que par construction (cf. les résultats de pas de gradient adaptatifs (Zinkevich, 2003) (Orabona, 2019)), le critère 2 soit vérifié. Le coût de l'adaptativité est uniquement une constante multiplicative $\sqrt{2}$ par rapport à la borne de regret que l'on obtient à $\gamma$ fixe. De plus, ce résultat est beaucoup plus fort que celui que l'on obtient à $\gamma$ fixe car il est valable à tout instant.

La principale nouveauté est inspirée de ce qui est fait par Gibbs and Candès (2022) et réside dans l'obtention d'une borne qui montre que l'algorithme 2 vérifie le critère 3 :

**Proposition 2.** *Soit* $\alpha_t^*$ *tel que* $\mathbb{P}\left(Y_t \in \hat{C}_t(\alpha_t^*) \mid \{\beta_s\}_{s<t}\right) = 1-\alpha$ *et* $p$ *un minorant positif de la densité de chacun des* $\beta_t$*. Les* $\alpha_t$ *construits par l'algorithme 2 vérifient :*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\left[(\alpha_t-\alpha_t^*)^2\right]}{2} \leq O\left(\frac{\sum_{t=2}^{T}\mathbb{E}\left[\left|\alpha_t^*-\alpha_{t-1}^*\right|\right]}{p\sqrt{T}}\right)$$

Cette proposition permet de mesurer à quel point les $\alpha_t$ sont proches des $\alpha_t^*$ en bornant le terme de gauche par un terme qui dépend uniquement des variations relatives des $\alpha_t^*$. On peut alors interpréter la borne comme étant un $O\left(\sqrt{T}\right)$ dans le pire des cas qui correspond à un environnement totalement adversarial avec la distribution des $Y_t$ qui change significativement à chaque instant. A l'inverse, si la distribution des $Y_t$ ne subit qu'un nombre fini de changements sur un horizon infini, on obtient donc une borne en $O\left(1/\sqrt{T}\right)$. On peut alors raisonnablement espérer qu'en pratique, nous sommes plutôt dans le second cas, ce qui rend cette proposition très intéressante pour les cas d'application.

**Remarque :** Cette seconde interprétation est, en fait, plus générale que ce qui est présenté par Angelopoulos et al. (2024). En effet, ils présentent pour leur méthode des garanties dans le cadre *i.i.d.* tandis que ce résultat permet également d'obtenir des résultats de convergence de l'estimateur dès lors que le nombre de changements de distribution est faible.

# 4 Comparaison

Dans cette section, il s'agit de comparer les garanties théoriques des différentes méthodes de type ACI. Parmi elles, on considère ACI, DtACI et surtout l'algorithme 2 pour lequel, nous avons montré de nouvelles garanties.

| Méthode | ACI | SF-OGD | DtACI | Algorithme 2 |
|---|---|---|---|---|
| $\gamma$ adaptatif | Non | Oui | Oui | Oui |
| Couverture 1 | $O\left(\frac{1}{T}\right)$ | $O\left(\frac{\log T}{T^{1/4}}\right)$ | $O(1)$ | $O\left(\frac{1}{T^{1/2}}\right)$ |
| Regret 2 | $DG\sqrt{T}$ | $(\sqrt{3}+1)DG\sqrt{t}$ | ? | $\sqrt{2}DG\sqrt{t}$ |
| Efficacité 3 | ? | ? | $\sqrt{\frac{\log T+\sum_{t=2}^{T}\mathbb{E}\left[\left\|\alpha_t^*-\alpha_{t-1}^*\right\|\right]}{T}}$ | $\frac{\sum_{t=2}^{T}\mathbb{E}\left[\left\|\alpha_t^*-\alpha_{t-1}^*\right\|\right]}{\sqrt{T}}$ |

Ce tableau permet de conclure que l'algorithme 2 offre parmi les meilleures garanties selon les critères que nous avons définis. Susmann et al. (2023) comparent également ces méthodes

en les appliquant sur des données simulées à l'aide de leur package R.

**Remarque :** Nous avons décidé de ne pas inclure SAOCP (Bhatnagar et al., 2023) dans le tableau comparatif car la nature des résultats et des hypothèses rendait difficile toute comparaison vis à vis des critères que nous avons énoncés.

# 5 Remerciements

Ce travail s'inscrit dans le cadre d'une thèse qui a débuté le 1er Décembre 2023 sous la supervision d'Yvenn Amara-Ouali, Bachir Hamrouche, Yannig Goude, Gilles Stoltz et Jean-Michel Poggi. Je les remercie pour les discussions ainsi que pour leur aide pour la rédaction.

# References

Amara-Ouali, Y., Hamrouche, B., Principato, G., and Goude, Y. (2024). Quantifying the uncertainty of electric vehicle charging with probabilistic load forecasting. In preparation.

Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024). Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. (2023). Improved online conformal prediction via strongly adaptive online learning. *arXiv preprint arXiv:2302.07869*.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.

Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*.

Orabona, F. (2019). A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*.

Susmann, H., Chambaz, A., and Josse, J. (2023). Adaptiveconformal: An R package for adaptive conformal inference. *arXiv preprint arXiv:2312.00448*.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936.

# Loi jointe et concentration des $p$-valeurs conformelles

Ulysse Gazin [1] & Gilles Blanchard [2] & Étienne Roquain[3]

[1] *Université Paris Cité and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation. Email: ugazin@lpsm.paris*
[2] *Université Paris Saclay, Institut Mathématique d'Orsay. Email: gilles.blanchard@universite-paris-saclay.fr*
[3] *Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation. Email: etienne.roquain@upmc.fr*

**Résumé.** L'inférence conformelle permet de construire des ensembles de prédiction en régression ou en classification et de faire de la détection de nouveautés via des méthodes issues de l'apprentissage. Cette approche permet de quantifier l'incertitude des procédures d'apprentissage. Les $p$-valeurs conformelles sont les outils de base de l'inférence conformelle, d'où l'importance d'une étude théorique de ces objets. Prendre $m$ décisions, ou bien construire $m$ régions de prédiction ou de confiance, demande de comprendre la loi des $p$-valeurs conformelles sous l'hypothèse classique d'échangeabilité. Même si les lois marginales sont connues, la loi jointe reste inconnue dans la littérature et on propose ici son étude. On présente la loi jointe des $p$-valeurs sous différents aspects: séquentiel avec les urnes de Pólya, bayésien avec une variable latente de Dirichlet, ainsi qu'explicite via une formule combinatoire. In fine, on montre une inégalité de concentration de type DKW pour la fonction de répartition empirique des $p$-valeurs conformelles. Ce résumé reprend l'article "Transductive conformal inference with adaptive scores" de U. Gazin, G. Blanchard et E. Roquain en se focalisant sur les propriétés théoriques des $p$-valeurs conformelles.

**Mots-clés.** Inégalité de concentration, $p$-valeur, Urne de Pólya, Inférence conformelle.

**Abstract.** Conformal inference have recently emerged as one of the fundamental tools in various domains of statistics by allowing to construct prediction sets for regression or classification and for machine learning novelty detection. This approach allows us to quantify the uncertainty of any machine learning procedure. Conformal $p$-values are the main tool of conformal inference and so a theoretical study is fundamental. Providing $m$ multiple decision (or simultaneous inference) in these contexts strongly relies on the joint distribution of the conformal $p$-values under an exchangeability assumption. However, while the marginal distribution is well known, only little is known about the joint distribution in the litterature and we propose to fill the gap. We present the joint distribution of conformal $p$-values under various aspects: a sequential one with a Pólya urn model, a Bayesian one with a latent Dirichlet random variable, and an explicit one with a combinatory formula. Finally, we propose a DKW-type concentration inequality for the empirical cumulative distribution function of $m$ conformal $p$-values. This report is a summary of the article "Transductive conformal inference with adaptive scores" by U. Gazin, G. Blanchard and E. Roquain with a focus on the theoretical properties of conformal $p$-values.

**Keywords.** Concentration inequality, $p$-value, Pólya urn model, Conformal inference.

# 1 Introduction

Les inégalités de concentration sont fondamentales en théorie des tests. Que ce soit pour la création de tests non asymptotiques lorsque l'on ne connaît pas explicitement la loi sous l'hypothèse nulle ou ses quantiles (pensons à l'inégalité DKW (Massart, 1990) pour le test non-asymptotique de Kolmogorov-Smirnov) ou pour contrôler un taux d'erreur en test multiple (Meah et al., 2023). Ici, nous cherchons à obtenir une propriété de concentration pour des variables aléatoires particulières: les $p$-valeurs conformelles, d'abord introduites par Saunders et al. (1999) pour l'inférence conformelle (conformal inference). Nous nous intéressons donc à l'étude de la fonction de répartition empirique de $m$ $p$-valeurs conformelles, que nous voyons ici comme un taux d'erreur, afin d'en déduire une inégalité de concentration de type DKW. Ce résumé est une relecture de l'article (Gazin et al., 2023) des mêmes auteurs; mais pris sous l'angle de l'étude théorique des $p$-valeurs conformelles. Nous renvoyons donc vers l'article en question pour voir plus en détails les preuves, les liens de ces $p$-valeurs avec les applications en prédiction conformelle et en détection de nouveautés ainsi que différents exemples d'applications en lien avec l'apprentissage, comme par exemple le transfer learning.

On pose, pour tout $(r < n) \in \mathbb{N}^2$, $[\![r, n]\!] := \{r, r+1, \cdots, n\}$, et $[\![n]\!] := [\![1, n]\!] = \{1, 2, \cdots, n\}$.

## 1.1 Les $p$-valeurs conformelles

On se donne deux familles de variables aléatoires réelles $\mathcal{D}_{\text{cal}} = \{S_1, \cdots, S_n\}$ et $\mathcal{D}_{\text{test}} = \{S_{n+1}, \cdots, S_{n+m}\}$ que nous appellerons scores et on définit comme suit nos $p$-valeurs conformelles.

**Définition 1.1** ($p$-valeurs conformelles)**.** Soit deux familles de variables aléatoires réelles $\mathcal{D}_{\text{cal}} = \{S_1, \cdots, S_n\}$ et $\mathcal{D}_{\text{test}} = \{S_{n+1}, \cdots, S_{n+m}\}$. On définit, pour tout $i \in [\![m]\!]$, la $i$-ième $p$-valeur conformelle $p_i$:

$$p_i = \frac{1}{n+1}\left(1 + \sum_{k \in [\![n]\!]} \mathbb{1}\{S_k \geq S_{n+i}\}\right). \tag{1}$$

On définit également la fonction de répartition empirique de ces $p$-valeurs conformelles:

$$\widehat{F}_m : t \in \mathbb{R} \mapsto \frac{1}{m} \sum_{i \in [\![m]\!]} \mathbb{1}\{p_i \leq t\} \in [0, 1]. \tag{2}$$

Si les variables aléatoires $(S_i)_{i \in [\![n+m]\!]}$ sont iid de loi $P_0$ de fonction de survie caglad $\overline{F}_- = \mathbb{P}(S_1 \geq \cdot)$, alors $p_i$ est proche de l'estimateur plug-in de $\overline{F}_-(S_{n+i})$ qui est une $p$-valeur (au sens de Lehmann and Romano (2005)) pour un test d'adéquation à la loi $P_0$ :

$$(\mathrm{H}_0) : \text{" La loi de } S_{n+i} \text{ est } P_0 \text{ "},$$
$$\text{contre } (\mathrm{H}_1) : \text{" La loi de } S_{n+i} \text{ n'est pas } P_0 \text{ "}.$$

En ce sens, les $p$-valeurs conformelles peuvent être vue comme des $p$-valeurs empiriques, et sont d'ailleurs utilisées par Bates et al. (2023) et Marandon et al. (2022) pour faire de la détection de nouveautés en appliquant une procédure de test multiples.

On introduit les deux hypothèses nécessaires pour l'étude de ces $p$-valeurs.

**Hypothèse 1.** *Les scores* $(S_1, \cdots, S_n, S_{n+1}, \cdots, S_{n+m})$ *sont échangeables.*

Cette hypothèse, un peu plus faible que i.i.d., provient de la littérature classique en inférence conformelle (on pourra lire Angelopoulos and Bates (2021) pour une introduction), mais se justifie par l'égalité suivante. En effet, $(n+1)p_i = \left|\{k \in [\![n]\!] \cup \{n+i\}, S_k > S_{n+i}\}\right|$ ce qui correspond, dans le cas où les $(S_j)_j$ sont p.s. deux à deux disjoints au rang de $S_{n+i}$ dans l'ensemble $\{S_1, \cdots, S_n, S_{n+i}\} = \left\{S_{(1)} > S_{(2)} > \cdots > S_{(n)} > S_{(n+1)}\right\}$ des scores ordonnés dans l'ordre décroissant. La loi des statistiques de rang étant libre de la loi de $(S_j)_j$ et connue dès que les scores sont échangeables et distincts p.s., l'hypothèse que les scores soient indépendants est superflue. L'hypothèse suivante complète la première afin d'utiliser cette loi des statistiques de rang.

**Hypothèse 2.** *Les scores* $(S_1, \cdots, S_n, S_{n+1}, \cdots, S_{n+m})$ *sont deux à deux distincts presque sûrement.*

Cette hypothèse est faible dans le sens où si les scores ne la vérifient pas il est toujours possible de les perturber en rajoutant un "petit" bruit indépendant pour avoir de nouveaux scores proches des premiers; mais vérifiant désormais l'hypothèse 2.

## 1.2 Lois marginales

Tout d'abord, les $p$-valeurs conformelles sont bien des $p$-valeurs car elles sont sur-uniformes.

**Théorème 1.2** (Romano and Wolf (2005))**.** *Supposons que l'hypothèse 1 soit vraie. Alors, pour tout* $i \in [\![m]\!]$, $p_i$ *est sur-uniforme, i.e.*

$$\forall \alpha \in [0,1],\ \mathbb{P}(p_i \leq \alpha) \leq \alpha.$$

De plus, on connaît la loi marginale si les hypothèses 1 et 2 sont vérifiées.

**Théorème 1.3.** *Supposons que les hypothèses 1 et 2 soient vérifiées. Alors, pour tout* $i \in [\![m]\!]$, $p_i$ *suit la loi uniforme sur* $[\![n+1]\!]/(n+1)$.

On peut remarquer, que du théorème 1.2, on peut construire un test non-asymptotique d'échangeabilité de $(S_1, \cdots, S_n, S_{n+1})$ avec la $p$-valeur $p_1$ en rejetant l'hypothèse nulle si $p_1 \leq \alpha$. En faisant cela pour tout $i \in [\![m]\!]$, $\widehat{F}_m(\alpha)$ correspond au nombre d'hypothèses nulles rejetées au niveau $\alpha$ divisé par le nombre $m$ de tests. De ces garanties marginales, on peut déjà obtenir un premier résultat sur la fonction de répartition empirique. On introduit d'abord, pour tout $n \in \mathbb{N}$, $I_n$ la fonction de répartition de la loi uniforme sur $[\![n+1]\!]/(n+1)$. On a, pour tout $t \in \mathbb{R}$, $I_n(t) = \mathbb{1}\{t \geq 1\} + \mathbb{1}\{t \in ]0,1[\}\frac{\lfloor (n+1)t \rfloor}{n+1}$.

**Corollaire 1.4.** *Si les hypothèses 1 et 2 sont vérifiées, alors pour tout* $t \in \mathbb{R}$, $\mathbb{E}\Big(\widehat{F}_m(t)\Big) = I_n(t)$.

Cependant, on ne peut en dire plus sur la loi jointe et sur $\widehat{F}_m$ sans étudier plus particulièrement la loi jointe des $(p_i)_{i\in[\![m]\!]}$. En effet, en regardant la définition des $p$-valeurs conformelles, on remarque dans (1) que les scores de calibration $S_1, \cdots, S_n$ sont présents dans tous les $(p_i)_{i\in[\![m]\!]}$, et de là on peut supposer que les $p$-valeurs conformelles ne sont pas indépendantes. Cette dépendance est un problème puisque l'on ne peut désormais plus utiliser les théorèmes usuels sur les variables aléatoires i.i.d. (Boucheron et al., 2013; Shorack and Wellner, 1986).

# 2 Loi jointe des $p$-valeurs et une inégalité de concentration de type DKW

## 2.1 $P_{n,m}$ : la loi jointe des $p$-valeurs conformelles

Comme expliqué ci dessus, les $p$-valeurs conformelles données par (1) ne sont pas indépendantes; mais on remarque qu'elle le sont conditionnellement à $\mathcal{D}_{\text{cal}}$. Ce résultat, classique dans la littérature (on le trouve par exemple dans la preuve du théorème 6 de Bates et al. (2023)), nous indique que nous manipulons des $p$-valeurs "quasiment" indépendantes, et qui le deviennent lorsque l'on conditionne par les bonnes variables aléatoires. On définit ici une loi $P_{n,m}$ particulière vérifiant de telles propriétés.

**Définition 2.1** ($P_{n,m}$)**.** Soit $n \geq 1$ et $m \geq 1$ deux entiers. Soit $U = (U_i)_{i\in[\![n]\!]}$ un échantillon de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$. On définit la distribution discrète $P^U$ sur l'ensemble $\left\{\frac{k}{n+1}, k \in [\![n+1]\!]\right\}$ comme suit:

$$\forall k \in [\![n+1]\!],\ P^U(\{k/(n+1)\}) = U_{(k)} - U_{(k-1)},$$

où $0 =: U_{(0)} < U_{(1)} < \cdots < U_{(n)} < U_{(n+1)} := 1$ sont les valeurs ordonnées de $U = (U_1, \cdots, U_n)$ dans l'ordre croissant. On définit désormais $(q_1, \cdots, q_m)$ $m$ variables aléatoires qui conditionnellement à $U$ sont i.i.d. de loi $P^U$. On note $P_{n,m}$ la loi inconditionnelle de $(q_1, \cdots, q_m)$. En résumé, $P_{n,m}$ est la loi sur $[0,1]^m$ définie comme suit:

$$P_{n,m} = \mathcal{D}\big(q_i, i \in [\![m]\!]\big),\ \text{où} \tag{3}$$

$$\begin{cases} \big(q_1, \ldots, q_m | U\big) & \overset{\text{i.i.d.}}{\sim} P^U; \\ \text{et } U = (U_1, \ldots, U_n) & \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1). \end{cases} \tag{4}$$

La mesure aléatoire $P^U$ est un processus de Dirichlet discret, dans le sens où le vecteur $\big(P^U\big(\frac{k}{n+1}\big)\big)_{k\in[\![n+1]\!]}$ suit une loi de Dirichlet de paramètre $(1, \cdots, 1) \in \mathbb{R}^{n+1}$.

Dans le cas où les scores sont i.i.d. et distincts p.s., il est montré dans la littérature que les $p$-valeurs conformelles ont pour loi $P_{n,m}$: les $(p_1, \cdots, p_m)$ sont i.i.d. conditionnellement à $\mathcal{D}_{\text{cal}}$ avec pour loi marginale $P^{U(\mathcal{D}_{\text{cal}})}$ où $U(\mathcal{D}_{\text{cal}}) = (1 - F(S_i))_{i\in[\![n]\!]}$ avec $F$ la fonction de répartition de $S_1$. De là, en remarquant que les $(p_i)_{i\in[\![m]\!]}$ se définissent à partir des statistiques de rang seules, on peut en déduire la loi jointe des $p$-valeurs conformelles en utilisant uniquement l'échangeabilité des scores $(S_i)_{i\in[\![n+m]\!]}$ (et non plus l'indépendance).

**Proposition 2.2** (Gazin et al. (2023))**.** *Si les hypothèses 1 et 2 sont vérifiées, alors les $p$-valeurs conformelles $(p_i)_{i\in[\![m]\!]}$ ont pour loi $P_{n,m}$ qui est donc libre de la loi des scores.*

L'universalité de $P_{n,m}$, dans le sens où la loi des $p$-valeurs conformelles est libre de la loi sous-jacente des scores n'est pas surprenante. En effet, en théorie des tests on cherche à construire des statistiques de test dont nous pouvons contrôler la loi sous $(\mathrm{H}_0)$, ce qui revient à trouver des statistiques de test libres de la loi sous $(\mathrm{H}_0)$. Par exemple, lors d'un test d'adéquation à une loi continue, il est connu que la loi de la statistique de test de Kolmogorov-Smirnov sous $(\mathrm{H}_0)$ est libre de la loi sous-jacente, et peut donc être simulée en utilisant des variables aléatoires uniformes. De la même façon, la proposition 2.2 montre que l'on peut simuler la loi jointe de $(p_i)_{i\in[\![m]\!]}$ à partir de variables aléatoires uniformes. La loi $P_{n,m}$ n'est pas une nouveauté en probabilités, et correspond à un phénomène bien connu: les $(p_1, \cdots, p_m)$ correspondent à $m$ tirages dans une urne de Pólya contenant $n+1$ boules de couleurs différentes. On définit pour tout $\boldsymbol{j} = (j_1, \ldots, j_m) \in [\![n+1]\!]^m$, le vecteur $\boldsymbol{M}(\boldsymbol{j}) := (M_k(\boldsymbol{j}))_{k\in[\![n+1]\!]}$ où pour tout $k \in [\![n+1]\!]$, $M_k(\boldsymbol{j}) := |\{i \in [\![m]\!] \,:\, j_i = k\}|$ est le nombre de coordonnées dans $\boldsymbol{j}$ égales à $k$ et dans ce cas on pose $\boldsymbol{M}(\boldsymbol{j})! := \prod_{k=1}^{n+1}(M_k(\boldsymbol{j})!)$.

**Théorème 2.3.** *La loi $P_{n,m}$ correspond à la distribution des couleurs de $m$ tirages successifs dans une urne de Pólya avec $n+1$ couleurs numérotées $\left\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\right\}$ et initialement une seule boule de chaque couleur. C'est-à-dire, si $\boldsymbol{p} \sim P_{n,m}$ définie en (3), on a les propriétés suivantes.*

*(i) Description dynamique : pour tout $i \in [\![0, m-1]\!]$, la loi de $p_{i+1}$ conditionnellement à $p_1, \ldots, p_i$ ne dépend pas de $m$ et est donnée par*

$$\mathcal{D}(p_{i+1} \mid p_1, \ldots, p_i) = \sum_{j=1}^{n+1} \frac{1 + \sum_{k=1}^{i} \mathbb{1}\{p_k = j/(n+1)\}}{n+1+i} \delta_{j/(n+1)}. \tag{5}$$

*(ii) Loi jointe : pour tout $\boldsymbol{j} \in [\![n+1]\!]^m$,*

$$\mathbb{P}\left(\boldsymbol{p} = \frac{\boldsymbol{j}}{n+1}\right) = \boldsymbol{M}(\boldsymbol{j})! \frac{n!}{(n+m)!}. \tag{6}$$

*(iii) Distribution de l'histogramme: l'histogramme de $\boldsymbol{p}$ est distribué uniformément sur l'ensemble des histogrammes d'un $m$-échantillon en $n+1$ subdivisions, i.e. pour tout $\boldsymbol{m} = (m_1, \ldots, m_{n+1}) \in [\![0, m]\!]^{n+1}$ tel que $m_1 + \cdots + m_{n+1} = m$, on a*

$$\mathbb{P}\big(\boldsymbol{M}\big((n+1)\boldsymbol{p}\big) = \boldsymbol{m}\big) = \binom{n+m}{m}^{-1}. \tag{7}$$

*En particulier, conditionnellement à $\boldsymbol{M}\big((n+1)\boldsymbol{p}\big)$, la variable $\boldsymbol{p}$ est uniformément distribuée sur l'ensemble des trajectoires possibles, c'est à dire, pour tout vecteur $\boldsymbol{j} \in [\![n+1]\!]^m$,*

$$\mathbb{P}\left(\boldsymbol{p} = \frac{\boldsymbol{j}}{n+1} \;\middle|\; \boldsymbol{M}\big((n+1)\boldsymbol{p}\big) = \boldsymbol{M}\big(\boldsymbol{j}\big)\right) = \frac{\boldsymbol{M}\big(\boldsymbol{j}\big)!}{m!}. \tag{8}$$

Un corollaire intéressant du point (*iii*) consiste à dire que $\widehat{F}_m$ suit une loi uniforme sur l'ensemble des fonctions de répartition des variables aléatoires à valeurs dans $[\![n+1]\!]/(n+1)$ faisant $m$ sauts, i.e., telle que la probabilité de valoir $k/(n+1)$ soit de la forme $\ell/m$ avec $\ell \in [\![0, m]\!]$ un entier, et cela pour tout $k \in [\![n+1]\!]$.

## 2.2 Concentration de la fonction de répartition empirique

La quantité $\widehat{F}_m(t)$ est intéressante car elle représente la proportion d'erreurs faites au niveau $t$ sur $m$ points de tests comme énoncé dans Gazin et al. (2023). En utilisant la proposition 2.2 avec le théorème 2.3 il est possible de retrouver la distribution de $\widehat{F}_m(t)$ à $t \in [0,1]$ fixé et nous permet de retrouver les résultats de Marques F. (2023). Cependant dans le cadre d'un contrôle d'un taux d'erreur il faut connaître le processus $(\widehat{F}_m(t))_{t\in[0,1]}$. À cette fin la proposition 2.2 nous indique qu'il existe une certaine variable aléatoire $U = (U_1, \cdots, U_n)$ telle que conditionnellement à $U$, $\widehat{F}_m$ soit la fonction de répartition empirique d'un $m$-échantillon de loi $P^U$. Cela nous permet donc d'appliquer les théorèmes usuels, comme l'inégalité DKW par exemple, lorsque l'on conditionne, puis d'intégrer contre le conditionnement en sachant que le vecteur $U$ est un $n$-échantillon de loi $\mathcal{U}(0,1)$. On peut ainsi obtenir l'inégalité de concentration de type DKW suivante.

**Théorème 2.4** (Gazin et al. (2023))**.** *Supposons que les hypothèses 1 et 2 soient vérifiées. Alors on a les inégalités suivantes. Pour tout $\lambda \geq 0$,*

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\in\mathbb{R}}\Big(\widehat{F}_m(t) - I_n(t)\Big) > \lambda\right) &\leq \mathbb{1}\{\lambda < 1\}\left[1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}}\lambda\right]e^{-2\tau_{n,m}\lambda^2};\\
\mathbb{P}\left(\sup_{t\in\mathbb{R}}\Big(\widehat{F}_m(t) - I_n(t)\Big) < -\lambda\right) &\leq \mathbb{1}\{\lambda \leq 1\}\left[1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}}\lambda\right]e^{-2\tau_{n,m}\lambda^2};\\
\mathbb{P}\Big(\Big\|\widehat{F}_m - I_n\Big\|_\infty > \lambda\Big) &\leq \mathbb{1}\{\lambda \leq 1\}\left[1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}}\lambda\right]2e^{-2\tau_{n,m}\lambda^2};
\end{aligned} \tag{9}$$

*avec $\tau_{n,m} = \frac{nm}{n+m} \in [n \wedge m/2, n \wedge m]$.*

Cette inégalité de concentration nous donne une vitesse de concentration de l'ordre de $\sqrt{\tau_{n,m}}$, qui est également de l'ordre de l'écart type, puisque l'on peut obtenir, si les hypothèses 1 et 2 sont vérifiées, que pour tout $(t,s) \in \mathbb{R}^2$,

$$\mathrm{Cov}\Big(\widehat{F}_m(t) - I_n(t), \widehat{F}_m(s) - I_n(s)\Big) = \frac{m+n+1}{m(n+2)}(I_n(t) \wedge I_n(s) - I_n(t)I_n(s)),$$

avec $\frac{m+n+1}{m(n+2)} \sim \tau_{n,m}$ quand $n \wedge m \to +\infty$. On peut remarquer, qu'en prenant le bon $\lambda$, on peut construire un test de niveau $\alpha$ testant si les scores sont échangeables en utilisant comme statistique de test $\left\|\widehat{F}_m - I_n\right\|_\infty$.

**Corollaire 2.5.** *Soit $\delta \in ]0,1[$. On définit, pour tout $r \in \mathbb{N}\backslash\{0\}$,*

$$\lambda^{DKW}_{\delta,n,m} = \Psi^{(r)}(1) \tag{10}$$
$$\text{où pour tout } x \in [0,1],\ \Psi(x) = 1 \wedge \left( \frac{\log\left(\frac{1}{\delta}\right) + \log\left(1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}} x\right)}{2\tau_{n,m}} \right)^{1/2},$$

*où $\Psi^{(r)}$ est la fonction $\Psi$ composée $r$ fois. Alors, on a:*

$$\mathbb{P}\left( \sup_{t\in\mathbb{R}} \left( \widehat{F}_m(t) - I_n(t) \right) > \lambda^{DKW}_{\delta,n,m} \right) \leq \delta;$$
$$\mathbb{P}\left( \sup_{t\in\mathbb{R}} \left( \widehat{F}_m(t) - I_n(t) \right) < -\lambda^{DKW}_{\delta,n,m} \right) \leq \delta;$$
$$\mathbb{P}\left( \left\| \widehat{F}_m - I_n \right\|_\infty > \lambda^{DKW}_{\delta/2,n,m} \right) \leq \delta.$$

La formule (10) pour $r = 1$ permet de comprendre facilement la dépendance en les paramètres $n$ et $m$ issus des données et $\delta$ fixé à priori. Par contre, afin d'obtenir un intervalle de confiance plus petit (ou une zone de rejet plus grande dans le cadre d'un test) tout en restant au niveau $1-\delta$, il convient de choisir un $r$ grand. Dans le cadre des simulations effectuées dans Gazin et al. (2023), le paramètre $r$ est choisi égal à 8.

L'uniformité des bornes (9) de ce théorème permet également de construire des bornes "post hoc" (Blanchard et al., 2020) dans le sens où l'on peut prendre n'importe quel niveau $\hat{\alpha} \coloneqq \hat{\alpha}(S_1, \cdots, S_{n+m})$ dépendant des données tout en ayant la garantie

$$\mathbb{P}\left( \widehat{F}_m(\hat{\alpha}) > I_n(\hat{\alpha}) + \lambda^{\mathrm{DKW}}_{\delta,n,m} \right) < \delta.$$

# 3 Conclusion

Nous avons ici présenté certaines propriétés théoriques des $p$-valeurs conformelles, et explicité leur loi jointe ainsi qu'une propriété de concentration de leur fonction de répartition empirique. Ces résultats théoriques permettent notamment de quantifier l'incertitude des tâches de machine learning dans lesquelles $\widehat{F}_m$ représente un taux d'erreur. Nous renvoyons la lectrice intéressée ou le lecteur intéressé aux articles Marandon et al. (2022) et Gazin et al. (2023) pour plus de détails.

# References


Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *Ann. Statist.*, 51(1):149–178.

Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist.*, 48(3):1281–1303.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Gazin, U., Blanchard, G., and Roquain, E. (2023). Transductive conformal inference with adaptive scores. *arXiv preprint arXiv:2310.18108*.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Adaptive novelty detection with false discovery rate guarantee. *To appear in Annals of Statistics*.

Marques F., P. C. (2023). On the universal distribution of the coverage in split conformal prediction. *arXiv preprint 2303.02770*.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283.

Meah, I., Blanchard, G., and Roquain, E. (2023). False discovery proportion envelopes with consistency.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108.

Saunders, C., Gammerman, A., and Vovk, V. (1999). Transduction with confidence and credibility. In *16th International Joint Conference on Artificial Intelligence (IJCAI 1999)*, pages 722–726.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

# Session internationale FENStatS/SFdS
# Statistique et Sport

# Improved handball match predictions via Statistically Enhanced Learning (SEL) and team strengths estimation

Florian Felice [1]

[1] *Département de Mathématiques, Université du Luxembourg, Luxembourg*
*florian.felice@uni.lu*

**Résumé.** Ce travail présente une nouvelle approche pour prédire les résultats des matchs de handball. Nous exploitons la puissance de l'apprentissage statistiquement amélioré (SEL) pour estimer la force des équipes. Notre modèle de machine learning augmenté par SEL surpasse les méthodes de l'état de l'art, atteignant une précision de 80%. Nous démontrons comment la force des équipes est dérivée statistiquement et intégrée sous forme de covariables informatives au modèle. En comparant différents modèles sur des données de clubs de handball féminin, nous révélons le rôle crucial des caractéristiques SEL dans la prédiction précise. De plus, les méthodes d'explicabilité dévoilent les facteurs clés influençant les buts marqués par les équipes, offrant des informations précieuses aux entraîneurs pour affiner leurs stratégies d'avant-match. Ce cadre s'étend bien au-delà du handball, présentant le SEL comme une méthode générale d'extraction de caractéristiques applicable à divers domaines de données et techniques d'apprentissage, avec le potentiel d'améliorer la prédiction dans diverses disciplines.

**Mots-clés.** Sport, handball, apprentissage automatique, prédiction, force

**Abstract.** This work presents a novel approach for predicting handball match outcomes. It leverages the power of Statistically Enhanced Learning (SEL) to estimate teams strengths. Our SEL-augmented machine learning model surpasses state-of-the-art methods, achieving an accuracy of 80%. We demonstrate how team strengths are statistically derived and integrated as informative covariates within the model. Comparing different models on female handball club data, we reveal the pivotal role of SEL features in accurate prediction. Further, explainability methods unveil key factors influencing team goal scoring, offering valuable insights for coaches to fine-tune strategies before matches. This framework extends far beyond handball, presenting SEL as a general feature extraction method applicable to various data domains and learning techniques, with the potential to enhance predictions across diverse disciplines.

**Keywords.** Sports analytics, handball, machine learning, prediction, strength

## Bibliographie

Felice, F. (2023), Ranking Handball Teams from Statistical Strength Estimation, *arXiv preprint* arXiv:2307.06754.

Felice, F. and Ley, C. (2023), Prediction of Handball Matches with Statistically Enhanced Learning via Estimated Team Strengths, *arXiv preprint*, arXiv:2307.11777.

# Basketball Data Science: questions and answers through statistical analysis

Marica Manisera*†[1] and Zuccolotto Paola*

[1]Università degli Studi di Brescia = University of Brescia – Italie

**Résumé**

The use of statistical methods and algorithms to find answers to sport questions is rapidly growing, thanks to an increasing availability of data and analytic tools. In this talk we provide some examples concerned with a selection of issues relevant in basketball: spatial performance measurement, evaluation of network relationships among players, assessment of injury risk factors.
For each example, we address the features of data, the exploited statistical methods, the useful information extracted, the open research issues.

**Mots-Clés:** basketball analytics, sports statistics, performance measurement, network analysis, injury risk factors

*Intervenant
†Auteur correspondant: marica.manisera@unibs.it

# Statistics meets Football

Gerharz Alexander*[1]

[1]Technische Universität Dortmund [Dortmund] – Allemagne

**Résumé**

In this talk, I will give an overview of different Data Sources and Application Fields of Statistics in a Football Club. This will include physical fitness data in training and matches. Also a lot of monitoring and diagnostics is done to get an overview about the players.

**Mots-Clés:** Football, Sports, Statistics

*Intervenant

# Données de survie, données censurées

# Statistical inference for the semi-parametric proportional reversed hazard model for left-censored and zero-inflated data[1]

Christian Paroissin [1] & Magdalena Pereda Vivo [2]

[1] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France et christian.paroissin@univ-pau.fr*

[2] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France et mpvivo@univ-pau.fr*

**Résumé.** Dans ce travail, on envisage d'analyser des données censurées à gauche avec une inflation de zéros. Cependant, il n'est pas possible de différencier une valeur zéro d'une observation positive censurée à gauche. Dans la littérature, certains articles proposent un modèle de mélange pour ce type de données, mais ils assument une distribution paramétrique pour les valeurs positives strictes. Ici, on considère un modèle de régression semi-paramétrique, plus précisément le modèle de risque inverse proportionnel, pour la partie positive. On propose ensuite un modèle de mélange semi-paramétrique pour traiter les données zero-inflatées censurées à gauche et on analyse l'influence des covariables sur les variables. En plus, on présente un algorithme EM pour estimer les différents paramètres et on étudie les propriétés asymptotiques des estimateurs. Cette méthodologie a été appliquée à des données simulées.

**Mots-clés.** Censure à gauche, excès de zéro, fonction de risque inverse

**Abstract.** In this work, we aim to analyse data subject to left censoring with inflation of zeros. It is not possible to distinguish a zero value from a positive left-censored observation. In the literature, there are some articles which propose a mixture model for this type of data but they assumed a parametric distribution for the strict positive values. Here, we consider a semi-parametric regression model, more precisely the proportional reversed hazard model, for the positive part. We then propose a semi-parametric mixture model for dealing with left-censored zero-inflated data and analyse the influence of the covariates on the variables. Furthermore, we provide an EM algorithm to estimate the different parameters and we study the asymptotic properties of the estimators. This methodology has been applied to simulated data.

**Keywords.** Left-censoring, zero excess, likelihood, EM algorithm

## 1 Introduction

In many areas (like toxicology, ecotoxicology, chemistry, geosciences and more generally in environmental sciences, as few examples), studies are based on data obtained by some an-

[1] *This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945416.*

alytical methods. However, with such approaches, one may have only partial information. For instance, when dealing with concentration measurements with an analytical method, one will observe an exact measurement only if it is larger than a certain threshold, called limit of quantification (LOQ): in other words, one has only the information that the concentration lies between zero and this limit. Such a situation is called left-censoring.

Censoring is well-known phenomena in statistics. It appears, for instance, in biostatistics. When dealing with lifetime data, some duration may not be observed exactly since the event occurs later than a certain time point. A typical case is when one performs a medical study over a given period: all the lifetimes longer than this period then get censored. Such a situation is known as right censoring, and it has been investigated rather extensively in the literature. Left censoring has been less studied by statisticians.

Besides, in some situations there could be the absence of the substance under consideration, and this will lead to true zeros (see Blackwood (1991), for example). In statistical terms, we talk about data with zero excess. We can relate the zero-inflated left-censored mixture models to the mixture cure models. In survival analysis, sometimes a part of the population could not experience the event of interest at the end of the follow-up period (susceptible individuals). Those data sets have right censoring. For such a situation, some authors have proposed a mixture cure model. Kuk and Chen (1992) proposed a semi-parametric mixture cure model using a Cox's model in the susceptible group and the logistic regression model for the cure fraction. Sy and Taylor (2000) developed maximum likelihood techniques for the estimation of the parameters in this model using the non-parametric form of the likelihood and an EM algorithm.

In this work, we aim to analyse data subject to left censoring with inflation of zeros. It is not possible to distinguish a zero value from a positive left-censored observation. In the literature, there are two articles which propose a mixture model for this type of data (Moulton (1995) and Yang (2010)), but they assumed a parametric distribution for the strict positive values. Here, we consider a semi-parametric regression model, more precisely the proportional reversed hazard model, for the positive part. We can find in Grouwels (2015) a semi-parametric regression model but they propose a Cox model for the positive part. We find more natural to use the reversed hazard function instead of the hazard function when dealing with left-censoring. For that, we propose a semi-parametric mixture model for dealing with left-censored zero-inflated data and analyse the influence of the covariates on the variables. Furthermore, we provide an EM algorithm to estimate the different parameters and we study the asymptotic properties of the estimators. This methodology has been applied to simulated data.

# 2 Notations

We assume that the observations are the realisations of independent and identically distributed random variables, conditionally to the covariates. We consider the case of multiple random censoring (type III left censoring scheme). Below, we introduce some notations that we will be used in the sequel.

- $n$ is the number of observations (exact or left-censored);
- $T_1, \ldots, T_n$ are the exact measurements (not always observed);
- $C_1, \ldots, C_n$ are the censoring values (not always observed) corresponding to LOQ;
- $X_1, \ldots, X_n$ are the observed values (exact or left-censored);
- $Z_1, \ldots, Z_n$ are the $p$-dimensional covariates;
- $\delta_1, \ldots, \delta_n$ are the indicators of the observation of exact values;
- $m$ is the number of distinct (exact or not) observations;
- $x_{(1)} < \cdots < x_{(m)}$ are the ordered distinct (exact or not) observations:

$$x_{(1)} = \min\{x_i\} < x_{(2)} < \cdots < x_{(m)} = \max\{x_i\};$$

- for any $k \in \{1, \ldots, m\}$, $d_k$ is the number of exact and observed measurements equal to $x_{(k)}$:

$$d_k = \#\{i \in \{1, \ldots, n\} : x_i = x_{(k)} \text{ and } \delta_i = 1\};$$

- for any $k \in \{1, \ldots, m\}$, $q_k$ is the number of left-censored and observed measurements equal to $x_{(k)}$:

$$q_k = \#\{i \in \{1, \ldots, n\} : x_i = x_{(k)} \text{ and } \delta_i = 0\};$$

- for any $k \in \{1, \ldots, m\}$, $y_k$ is the number of observations less than or equal to $x_{(k)}$:

$$y_k = \#\{i \in \{1, \ldots, n\} : x_i \leqslant x_{(k)}\} = \sum_{j=1}^{k} (d_j + q_j);$$

We assume that $(X_1, \Delta_1, Z_1), \ldots, (X_n, \Delta_n, Z_n)$ is an i.i.d. sample of the observed variables $(X, \delta, Z)$.

# 3 Proportional reversed hazard model with inflation of zeros

We denote by $T$ a semi-continuous variable which can have the value zero or positive values (non-zero values). In addition, we assume that some covariates $Z$ may have influence both on the probability to have a zero and on the distribution for the continuous part. More precisely, we consider the following semi-parametric mixture model for the variable $T$ :

$$F_{T|Z}(t|z) = P(T \leq t|Z = z) = \pi(b, z) + (1 - \pi(b, z))F_{T>0|Z}(t|z) \tag{1}$$

where $F_{T>0}(t|z) = P(T \leq t|T > 0, Z = z)$ is the continuous conditional distribution for the non-zero values of $T$ and $\pi(b, z) = P(T = 0|Z = z)$ is the conditional probability of a zero outcome value. We assume a parametric model for $\pi(b, z)$ where $b$ is the vector of parameters.

For the conditional distribution of the non-zero values, we consider a proportional reversed hazard model. When dealing with possibly left censored data, the reversed hazard function is more natural than the hazard function which is more suitable for the situation of right censoring. As studied by Sengupta and Nanda [**?**], this regression model proposed has the following form:

$$r_{T>0|Z}(t|z) = \frac{f_{T>0|Z}(t|z)}{F_{T>0|Z}(t|z)} = r(t)g(\beta, z), \tag{2}$$

where $r(t)$ is the baseline reversed hazard function and $g(\beta, z)$ is a non-negative function. As for the so-called Cox model, a natural choice for $g$ is the exponential function and the linear relationship for $\beta$ and $z$. In such a case, we have :

$$g(\beta, z) = \exp(z'\beta) = \prod_{j=1}^{p} \beta_j z_j.$$

Let $R(t|Z)$ be the cumulative reversed hazard function of $T$ given $Z$. Taking into account the relation between the reserved hazard function and the cumulative distribution function, we have that

$$F_{T>0|Z}(t|z) = \exp\left\{-\int_t^\infty r_{T>0}(s|z)ds\right\} = \exp\{-g(\beta, z)R(t)\}$$

where $R(t) = \int_t^\infty r(s)ds$ is the baseline cumulative reversed hazard function, which is equivalent to say that

$$F_{T>0|Z}(t|z) = [F(t)]^{g(\beta,z)}$$

where $F(t)$ is the baseline cumulative distribution function.

For this semi-parametric model, the different parameters are the following ones: $b$, $\beta$ and $F$ (or equivalently $r$). The two first ones are Euclidean parameters while the third one is a functional parameter.

For now, we will consider the estimation of the parameters by considering the case of type III left censoring (random censoring). We assume that there exists a random variable $C$ such that we only observe $X = \max(T, C)$ and $\Delta = I(T \geq C)$. Conditionally on the covariate $Z$, we assume that $T$ and $C$ are independent. The observations will be thus $(X_i, \Delta_i, Z_i)$ for $i = 1, \ldots, n$.

## 3.1 Estimation with EM algorithm

In this section, we estimate the parameters $\beta$ and $\gamma$ which maximize $L(\gamma, \beta, F)$. To do so, we will use the Expectation-Maximization (EM) algorithm proposed in Dempster (1997). This is an approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. The missing part is due to situation that, if an observation is smaller the limit of quantification (LOQ), we have not the information if this is a zero or a positive value between zero and the LOQ.

The observed full likelihood for the previous model is

$$L(b,\beta,F)=\prod_{i=1}^{n}[(1-\pi(b,z_i))f_{T>0}(X_i|Z_i)]^{\delta_i}[\pi(b,Z_i)+(1-\pi(b,z_i))F_{T>0}(X_i|Z_i)]^{1-\delta_i}.$$

We define $\widetilde{\Delta}_i=1$ if $T_i>0$ and $\widetilde{\Delta}_i=0$ if $T_i=0$, which is not always observed. Denote the complete data by $(X_i,\Delta_i,Z_i,\widetilde{\Delta}_i)$, which includes the observed data and unobserved $\widetilde{\Delta}_i$. If $\delta_i=1$ we know the exact value of $T_i>C_i>0$, then $\widetilde{\delta}_i=1$. Otherwise, $\widetilde{\delta}_i$ is unobserved since $T_i$ is lower than the censored threshold and we do not know if it is positive. The complete-data full likelihood is

$$L_C(b,\beta,F;\widetilde{\delta}) \;=\; \prod_{i=1}^{n}[\pi(b,Z_i)]^{1-\widetilde{\delta}_i}[(1-\pi(b,Z_i))]^{\widetilde{\delta}_i}r_{T>0}(X_i|Z_i)^{\delta_i\widetilde{\delta}_i}F_{T>0}(X_i|Z_i)^{\widetilde{\delta}_i}.$$

Then, the complete data full log-likelihood is

$$\ell_C(b,\beta,F;\widetilde{\delta})=\sum_{i=1}^{n}(1-\widetilde{\delta}_i)\log\pi(b,Z_i)+\widetilde{\delta}_i\log(1-\pi(b,Z_i))+\widetilde{\delta}_i\left[\delta_i\log r_{T>0}(X_i|Z_i)+\log F_{T>0}(X_i|Z_i)\right].$$

We choose an estimation approach based on the Expectation-Maximization algorithm because of the presence of the latent variable $\widetilde{\delta}$. The EM algorithm is based on iterative maximizations of the expectation of the log-likelihood relative to the complete data.

The E-step takes the expectation of $\ell_C(\gamma,\beta,F;\widetilde{\delta})$ with respect to the unobserved $\widetilde{\delta}_i$, which has the following form

$$\begin{aligned}\tilde{\ell}_C(b,\beta,R;\gamma(\widetilde{\delta})) &= \sum_{i=1}^{n}\Big\{[1-\gamma(\widetilde{\delta}_i)]\log\pi(b,Z_i)+\gamma(\widetilde{\delta}_i)\log(1-\pi(b,Z_i))\Big\}\\ &\quad+\sum_{i=1}^{n}\gamma(\widetilde{\delta}_i)[\delta_i\log r_{T>0}(X_i|Z_i)+\log F_{T>0}(X_i|Z_i)]\\ &= \widetilde{\ell}_1(b;\gamma(\widetilde{\delta}))+\widetilde{\ell}_2(\beta,R;\gamma(\widetilde{\delta})).\end{aligned}\tag{3}$$

where

$$\gamma(\widetilde{\delta}_i)=E[\widetilde{\delta}_i|X_i,\Delta_i,Z_i].$$

- If $\delta_i=0$, then $\widetilde{\delta}_i$ is unobserved and

$$\begin{aligned}\mathbb{E}(\widetilde{\delta}_i|X_i,\Delta_i,Z_i) &= \frac{P(\widetilde{\delta}_i=1,\delta_i=0,|X_i,Z_i)}{P(\delta_i=0,|X_i,Z_i)}=\frac{P(\widetilde{\delta}_i=1,|X_i,Z_i)P(\delta_i=0|\widetilde{\delta}_i=1)}{P(\delta_i=0,|X_i,Z_i)}\\ &= \frac{P(T_i>0)P(\delta_i=0|T_i>0)}{P(\delta_i=0)}=\frac{(1-\pi(\gamma,Z_i))F_{T>0}(X_i|Z_i)}{\pi(\gamma,Z_i)+(1-\pi(\gamma,Z_i))F_{T>0}(X_i|Z_i)}\end{aligned}\tag{4}$$

- If $\delta_i=1$, as $\widetilde{\delta}_i=1$, we get $\mathbb{E}(\widetilde{\delta}_i|X_i,\Delta_i,Z_i)=1$ .

In the M-step, we have to maximize $\tilde{\ell}_C(b, \beta, R; \gamma(\widetilde{\delta}))$ with respect to $b$, $\beta$ and $R$, given $\gamma(\widetilde{\delta})$. Since

$$\tilde{\ell}_C(b, \beta, R; \gamma(\widetilde{\delta})) = \widetilde{\ell}_1(b; \gamma(\widetilde{\delta})) + \widetilde{\ell}_2(\beta, R; \gamma(\widetilde{\delta}))$$

we can estimate $b$ and $\{\beta, R\}$ separately. We will combine EM algorithm and profile likelihood approach, since we need an estimate of the distribution function $F_{T>0}(X_i|Z_i)$ for calculate $\gamma(\widetilde{\delta}_i) = \mathbb{E}[\widetilde{\Delta}_i|X_i, \Delta_i, Z_i]$ in (4) and start using the EM algorithm.

First, we have to make an initial guess of the values of the parameters. We set $\widehat{b}^{(0)} = 0$ and we estimate $\widehat{R}_0^{(0)}$ assuming that $\widehat{\beta}^{(0)} = 0$ and using the non-parametric estimator of the cumulative reversed hazard function

$$\forall t \geqslant 0, \quad \widehat{R}^{(0)}(t) = \sum_{j; x_{(j)} > t} \frac{d_j}{y_j - q_j}.$$

In the $k$-th iteration of the EM algorithm, we proceed as follows.

1. In the E-step, we compute the expected values of $\widetilde{\delta}_i$ using the estimations $\widehat{b}^{(k-1)}$ and $\widehat{\beta}^{(k-1)}$ at the previous iteration:

$$\gamma(\widetilde{\delta}_i)^{(k)} = \delta_i + (1 - \delta_i) \frac{(1 - \pi(\widehat{b}^{(k-1)}, Z_i) \widehat{F}_{T>0}^{(k-1)}(X_i|Z_i; \widehat{\beta}^{(k-1)})}{\pi(\widehat{b}^{(k-1)}, Z_i) + (1 - \pi(\widehat{b}^{(k-1)}, Z_i)) \widehat{F}_{T>0}^{(k-1)}(X_i|Z_i; \widehat{\beta}^{(k-1)})},$$

taking into account that

$$\widehat{F}_{T>0}^{(k-1)}(X_i|Z_i; \widehat{\beta}^{(k-1)}) = \exp\{-g(\widehat{\beta}^{(k-1)}, Z_i) \widehat{R}^{(k-1)}(X_i; \widehat{\beta}^{(k-1)})\}$$

where

$$\widehat{R}^{(k-1)}(X_i; \widehat{\beta}^{(k-1)}) = \sum_{j=1}^{m} \left( \frac{d_j I(X_{(j)} > X_i)}{\sum_{l=1}^{n} \gamma(\widetilde{\delta}_l)^{(k-1)} I(X_{(j)} > X_l) g(\widehat{\beta}^{(k-1)}, Z_l)} \right), \ i = 1, \ldots, n.$$

2. The M-step consists in maximizing the expected likelihood with respect to the parameters of the model.

# Bibliographie


L G. Blackwood (1991), Analyzing censored environmental data using survival analysis: Single sample techniques, *Environmental Monitoring and Assessment*, 18, pp. 25-40.

Anthony Y. C. Kuk and Chen-Hsin Chen (1992), A mixture model combining logistic regression with proportional hazards regression, *Biometrika*, 79, pp. 531-541.

Sy, Judy P. and Taylor, Jeremy M. G. (2000), Estimation in a Cox Proportional Hazards Cure Model, *Biometric*, 56, pp. 227-236.

Lawrence H. Moulton and Neal A. Halsey (1995), A Mixture Model with Detection Limits for Regression Analyses of Antibody Response to Vaccine, *Wiley, International Biometric Society*, 51, pp. 1570-1578.

Yang, Yan and Simpson, Douglas (2010), Unified computational methods for regression analysis of zero-inflated and bound-inflated data, *Computational Statistics & Data Analysis*, 54, pp. 1525-1534.

Braekers, Roel and Grouwels, Yves (2015), A semi-parametric Cox's regression model for zero-inflated left-censored time to event data, *Communications in Statistics - Theory and Methods*, 45.

Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1997), Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, pp. 1-22.

Hirose, Yuichi and Liu, Ivy (2020), Statistical Generalized Derivative Applied to the Profile Likelihood Estimation in a Mixture of Semiparametric Models, *Entropy*, 22.

# Non-parametric estimation of net survival under dependence between death causes

Oskar Laverny [1] & Nathalie Graffféo [1] & Roch Giorgi [2]

[1] *Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.*
*oskar.laverny@univ-amu.fr*
[2] *Aix Marseille Univ, APHM, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Hop Timone, BioSTIC, Biostatistique et Technologies de l'Information et de la Communication, Marseille, France.*

**Résumé.** L'analyse de survie relative traite un problème de risques compétitifs où la cause de la mort est inconnue. Ce manque d'information arrive régulièrement dans les études de cohortes liées au cancer. L'estimation non paramétrique de la survie nette est possible via l'estimateur de Pohar Perme, qui prend en compte les autres causes de mortalité. Dérivé de manière analogue à Kaplan-Meier, cet estimateur repose cependant sur une hypothèse d'indépendance non testable. Nous proposons ici de relâcher cette hypothèse et fournissons un estimateur généralisé qui fonctionne sous d'autres structures de dépendances, en exploitant les processus de comptage et les martingales sous-jacentes. Notre approche fournit de nouvelles perspectives sur l'estimateur de Pohar Perme et l'acceptabilité de ses hypothèses. Nous montrons la différence entre les deux estimateurs sur des données relatives au cancer colorectal et discutons finalement de possibles extensions de la méthodologie.

**Mots-clés.** Analyse de survie, Surive nette, Estimateurs non-paramétriques, Copules.

**Abstract.** Relative survival analysis deals with a competing risks survival model where the cause of death is unknown. This lack of information occurs regularly in population-based cancer studies. Non-parametric estimation of the net survival is possible through the Pohar Perme estimator, taking other causes of mortality into account. Derived similarly to Kaplan-Meier, it nevertheless relies on untestable independence assumptions. We propose here to relax these assumptions and provide a generalized estimator that works for other dependence structures, by leveraging the underlying counting process and martingales. Our approach provides a new perspective on the Pohar Perme estimator and the acceptability of this assumption. We showcase the difference between the two estimators on population-based colorectal cancer registry, and discuss potential extensions of the methodology.

**Keywords.** Survival analysis; Net survival; Non-parametric estimators; Copulas

## 1 Net survival analysis

Survival analysis produces valuable tools for prognosis of cancer patients. However, in population-based cancer studies, the cause of death – assumed binary, studied cancer or

not – is usually unreliable or unavailable. Relative survival analysis takes this particularity into account to evaluate the excess mortality – due to cancer – with respect to population life tables.

Let $E, P$ and $O = E \wedge P$ be random times to death from (resp) the Excess, Population and Overall mortalities. Let $\mathbf{X}$ be a vector of covariates, $C$ the time to censorship, and denote $T = O \wedge C$ and $\Delta = \mathbf{1}\{T \leq C\}$. Only $(\mathbf{X}, T, \Delta)$ is observable. In particular, we do not observe a potential ordering indicatrix $\mathbf{1}\{E \geq P\}$. Alike standard approaches, we suppose the distribution of $P|\mathbf{X}$ known from population life tables. To simplify our exposition, assume that the rest does not depend on covariates.

Let $(\mathbf{X}_i, T_i, \Delta_i)_{i=1,..,n}$ be an observed $n$-sample and $(\Omega, \mathcal{A}, \{\mathcal{F}_t, t \in \mathbb{R}_+\}, \mathbb{P})$ the associated filtered probability space, with

$$\mathcal{F}_t = \sigma\{\mathbf{X}_i, (T_i, \Delta_i) : T_i \geq t,\ \forall i \in 1, .., n\}.$$

The mutual independence of $(E, P, C)$ is a central assumption in survival literature (see e.g., Czado & Van Keilegom (2023) for recent discussion), even if the epidemiological interpretation makes it hard to justify. We here suppose only independent censorship, which, denoting $\mathcal{C}$ the survival copula (see Nelsen (2006)) of the random vector $(E, P)$, writes:

$$(\mathcal{H}_{\mathcal{C}}):\ S_{P\wedge E}(t) = \mathcal{C}(S_P(t), S_E(t)).$$

In particular, the standard independence assumption writes simply $(\mathcal{H}_\Pi)$ for $\Pi(u_1, u_2) = u_1 u_2$. There are a few standard non-parametric estimators under $(\mathcal{H}_\Pi)$, from Ederer (1961), Hakulinen (1982), to more recently Pohar Perme & al (2012), but none under $(\mathcal{H}_{\mathcal{C}})$. If Adatorwovor & al (2023) provides parametric estimations under $(\mathcal{H}_{\mathcal{C}})$, we propose here a non-parametric estimator that generalizes the Pohar Perme estimator from $(\mathcal{H}_\Pi)$ to $(\mathcal{H}_{\mathcal{C}})$.

# 2 Estimation of the net survival under $(\mathcal{H}_{\mathcal{C}})$

Let's define the following stochastic processes:

$$\begin{aligned}
N(t) &= \mathbf{1}\{O \leq t, O \leq C\} \quad \textit{(Uncensored deaths process)}\\
Y(t) &= \mathbf{1}\{O \geq t, C \geq t\} \quad \textit{(At-risk process)}\\
N_E(t) &= \mathbf{1}\{E \leq t, E \leq C\} \quad \textit{(Excess uncensored deaths process)}\\
Y_E(t) &= \mathbf{1}\{E \geq t, C \geq t\} \quad \textit{(Excess at-risk process)}
\end{aligned}$$

Unfortunately, $(N_{E_i}, Y_{E_i})_{i\in1,...,n}$ are not observable, but we show that

$$\begin{aligned}
\partial N_E(t) &= \frac{1}{a_t}\mathbb{E}(\partial N(t)|E, C) - \frac{b_t}{a_t c_t}\mathbb{E}(Y(t)|E, C),\\
Y_E(t) &= \frac{1}{c_t}\mathbb{E}(Y(t)|E, C),
\end{aligned}$$

where $a_t = \mathbb{P}(P \geq t|E = t)$, $b_t = \mathbb{P}(P = t|E \geq t)$ and $c_t = \mathbb{P}(P \geq t|E \geq t)$.

Assuming that $(P, E)$ is an absolutely continuous random vector, $a_t, b_t, c_t$ can be computed from partial derivatives $\mathcal{C}_i(\mathbf{u}) = \frac{\partial \mathcal{C}}{\partial u_i}(\mathbf{u}), i \in 1, 2$ of $\mathcal{C}$. Remark that under $(\mathcal{H}_\Pi)$, $a_t = c_t = S_P(t)$ and $b_t = -\partial S_P(t)$ do not depend on (unknown) $S_E(t)$, while they generally do under $(\mathcal{H}_\mathcal{C})$. However, like in classical survival analysis, we show under $(\mathcal{H}_\mathcal{C})$ that $\mathbb{E}(\partial N_E(t)) = \mathbb{E}(Y_E(t)\partial\Lambda_E(t))$, and moreover that $N_{E_i} - \int_0^t Y_{E_i}\partial\Lambda_{E_i}$ are $\mathcal{F}_t$-martingales. A natural estimator for $\partial\Lambda_E(t)$ can therefore be simply constructed as

$$\widehat{\partial\Lambda_E}(t) = \frac{\frac{1}{n}\sum_{i=1}^n \partial N_{E_i}(t)}{\frac{1}{n}\sum_{i=1}^n Y_{E_i}(t)}.$$

However, $(\partial N_{E_i}, Y_{E_i})$ are not directly observable and need to be estimated. For that, replace first unobservable conditional expectations by their stochastic counterpart: $\mathbb{E}(\partial N_i(t)|E_i, C_i)$ by $\partial N_i(t)$ and $\mathbb{E}(Y_i(t)|E_i, C_i)$ by $Y_i(t)$. This plug-in is enough to make the estimator computable under $(\mathcal{H}_\Pi)$ (it is the Pohar Perme estimator), but not under $(\mathcal{H}_\mathcal{C})$. We call *generalized Pohar Perme estimator* a solution of the differential equation:

$$\widehat{\partial\Lambda_E}(t) = \frac{\sum_{i=1}^n \frac{1}{\widehat{a_{i,t}}}\partial N_i(t) - \frac{\widehat{b_{i,t}}}{\widehat{a_{i,t}}\widehat{c_{i,t}}}Y_i(t)}{\sum_{i=1}^n \frac{1}{\widehat{c_{i,t}}}Y_i(t)}, \tag{1}$$

where for all $i \in 1, ..., n$,

$$\begin{aligned}
\widehat{a_{i,t}} &= \mathcal{C}_2\left(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)}\right), \\
\widehat{b_{i,t}} &= -\mathcal{C}_1\left(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)}\right)\partial S_{P_i}(t)e^{\widehat{\Lambda_E}(t)}, \\
\widehat{c_{i,t}} &= \mathcal{C}(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)})e^{\widehat{\Lambda_E}(t)}.
\end{aligned}$$

Unfortunately, the differential equation 1 is now non-separable, and a non-linear equation in $\widehat{\partial\Lambda_E}(t)$ needs to be solved at each time step. Alike previous estimators under $(\mathcal{H}_\Pi)$, the obtained $\widehat{\partial\Lambda_E}$ process is piecewise continuous, with jumps at event times $T_1, ..., T_n$. It is moreover always negative, except at jump points, making the produced survival curve increasing between jumps. The solving scheme must therefore be performed on a very dense mesh $t_1, ..., t_N$ that includes observed times $T_1, ...T_n$. These characteristics were already present under $(\mathcal{H}_\Pi)$.

# 3 Illustration

We use data on colorectal cancer patients extensively described in Wolski & al (2020). Population is separated along the primary tumor location (left or right). Using several dependence structures, we obtain net survival curves from Figure 1. If Wolski & al (2020) found the overall survival to be significantly different between left and right, net survival might not be if we take into account the uncertainty in $\mathcal{C}$. Further analysis to derive proper log-rank-type tests under $(\mathcal{H}_\mathcal{C})$ is possible.

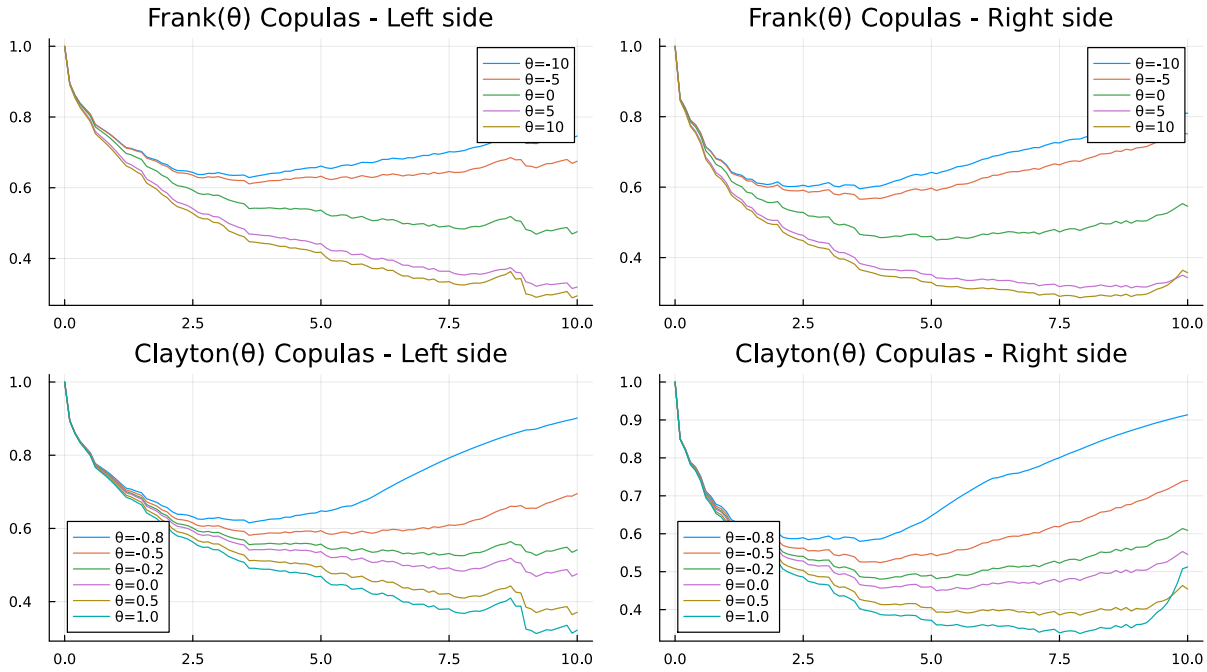



Figure 1: Obtained $\widehat{S_E}$ for several ($\mathcal{H}_{\mathcal{C}}$). Data was split w.r.t. tumor location (left or right) as in Wolski & al (2020), and several runs were done on several copulas: Frank copulas on the top and Clayton copulas on the bottom, with varying parameters $\theta$. In both cases, $\theta = 0 \iff \mathcal{C} = \Pi$, and this curve represents the Pohar Perme estimator.

# Bibliographie

**Adatorwovor, R., Latouche, A., and Fine, J. P.** (2023). A parametric approach to relaxing the independence assumption in relative survival analysis. In: *The international journal of biostatistics*, 18(2), 577-592.

**Czado, C. and Van Keilegom, I.** (2023). Dependent censoring based on copulas. In: *Biometrika* 110(3), 721-738.

**Ederer, F.** (1961). The relative survival rate: a statistical methodology. In: *Natl. Cancer Inst. Monogr.*, vol. 6, p. 101-121, 1961.

**Hakulinen, T.** (1982). Cancer survival corrected for heterogeneity in patient withdrawal. In : *Biometrics* 38, 933–942.

**Nelsen, R. B.** (2006). *An introduction to copulas*. Springer.

**Pohar Perme, M., Stare, J., and Estève, J.** (2012). On estimation in relative survival. In: *Biometrics*, vol. 68, no 1, p. 113-120

**Wolski, A., Grafféo, N., Giorgi, R., and the CENSUR working survival group** (2020). A permutation test based on the restricted mean survival time for comparison of net

survival distributions in non-proportional excess hazard settings. In: *Stat Methods Med Res*, vol. 29, no 6, p. 1612-1623

# Bayesian analysis of restricted mean survival time adjusted on covariates using pseudo-observations

Léa Orsini[1,2], Emmanuel Lesaffre[2], Guosheng Yin[3], Caroline Brard[4], David Dejardin[5] & Gwénaël Le Teuff[1]

[1] *CESP, INSERM U1018, Université Paris-Saclay, UVSQ, Villejuif, France, lea.orsini@gustaveroussy.fr, gwenael.leteuff@gustaveroussy.fr*
[2] *I-Biostat, KU-Leuven, Leuven, Belgium, emmanuel.lesaffre@kuleuven.be*
[3] *Department of Mathematics, Imperial College London, London, England, guosheng.yin@imperial.ac.uk*
[4] *Ipsen Innovation, Clinical Development Organisation, Les Ulis, France, caroline.brard@ipsen.com*
[5] *Product Development, Data Sciences, F. Hoffmann-La Roche AG, Basel, Switzerland, david.dejardin@roche.com*

**Résumé.** La différence de moyennes de temps de survie restreintes (dRMST) à un temps donné est une mesure adéquate pour quantifier l'effet traitement entre deux bras d'un essai clinique randomisé lorsque l'hypothèse des risques proportionnels n'est pas vérifiée. C'est une situation courante dans les essais d'immuno-oncologie. Plusieurs méthodes fréquentistes existent pour estimer la RMST, basées sur la modélisation et l'intégration de la fonction de survie. Une approche plus naturelle est de considérer un modèle de régression directement sur la RMST utilisant les pseudo-observations qui permettent d'éviter la modélisation de la fonction de survie. Cette approche offre aussi la possibilité d'étendre l'analyse de la RMST pour un temps donné à l'analyse jointe de la RMST à plusieurs temps. Seules deux méthodes Bayésiennes existent et modélisent la fonction de survie avec un mécanisme de priors non-paramétrique. Nous avons développé une nouvelle méthode Bayésienne, basée sur les pseudo-observations et la méthode des moments généralisées, qui offre une estimation de la RMST ajustée sur des covariables, sans avoir besoin de modéliser la fonction de survie, ce qui la rend attractive par rapport aux méthodes Bayésiennes existantes. Une étude de simulation d'essais randomisés à deux bras, avec différents effets traitement dépendant du temps et covariables, a été effectuée et montre que cette nouvelle méthode donne des résultats valides, cohérents avec les méthodes existantes et aussi une estimation plus précise après ajustement sur les covariables. A titre d'illustration, les différentes méthodes ont été appliquées aux données de l'essai Getug-AFU 15, essai randomisé de phase 3 comparant une thérapie de déprivation androgénique seule ou en complément du docetaxel chez des patients avec un cancer prostatique métastatique hormono-naif, pour analyser la survie sans progression biologique. Cette illustration montre également l'avantage de la nouvelle approche proposée pour analyser la RMST avec ajustement sur les covariables dans le cadre Bayésien.

**Mots-clés.** Méthode des Moments Généralisée Bayésienne, Risques non proportionnels, Pseudo-observations, Moyenne Restreinte du Temps de Survie.

**Abstract.** The difference in restricted mean survival time (dRMST) at a specific time point is an appropriate measure to quantify the treatment effect between two arms in randomized clinical trials when the proportional hazards assumption does not hold. This is common in the context of immuno-oncology therapies. Several frequentist methods exist to estimate RMST based on modeling and integrating the survival function. A more natural approach is to consider a regression model on the RMST directly using pseudo-observations which allows for a direct fit without modeling the survival function. This approach also has the advantage of extending the analysis for single time point of interest to the joint analysis at multiple time points. Only two Bayesian methods exist, and both model the survival function with a nonparametric prior process. We develop a new Bayesian method based on pseudo-observations and the generalized method of moments (GMM) that offers RMST estimation adjusted on covariates without the need to model the survival function, making it attractive compared to existing Bayesian methods. A simulation study of 2-arms randomized clinical trials with different time-dependent treatment effects and covariates effects was conducted, demonstrating that this new approach yields valid results, consistent with existing methods, and shows improved precision after covariates adjustment. For illustration, the methods were applied for analyzing the PSA progression-free survival of the Getug-AFU 15 trial, a randomized, open-label, phase 3 trial comparing an androgen-deprivation therapy alone or with docetaxel in non-castrate metastatic prostate cancer. This illustration also demonstrates the advantage of this new method to perform RMST analysis with covariates adjustment in the Bayesian framework.

**Keywords.** Bayesian generalized method of moments, Non-proportional hazards, Pseudo-observations, Restricted mean survival time, Survival Analysis.

# 1 Introduction

The difference of restricted mean survival time (dRMST) has been proposed as a clinically meaningful estimand of the treatment effect in randomized clinical trials, especially when the proportional hazards (PH) assumption does not hold. This situation often occurs in immuno-oncology therapies where the treatment effect is time-dependent (e.g., early or delayed effect). In this case, the interpretation of the hazard ratio as a treatment effect measure becomes difficult. Alternatively, the RMST can be used, interpretable as the expected life time experienced out of $\tau$ units of time. For example, if the dRMST between experimental and control groups at $\tau = 5$ years is 1 year, it means that, on average, the experimental treatment increases life expectancy during the next 5 years by 1 year, compared to the control treatment.

One straightforward approach to estimating RMST at a certain time $\tau$ is to numerically integrate the Kaplan-Meier curve between 0 and $\tau$. However, this approach does not allow for covariates adjustment, which is a major limitation because omitting important covariates results in less precision, see Karrison (2018). One way to adjust the RMST estimation on covariates is to model the survival function with a parametric or semi-parametric model and integrate it, see Karrison (1987) and Zucker (1998). A more natural approach is to fit a

regression model on the RMST directly through estimating equations instead of modeling the survival function. In this case, censoring must be handled either using the inverse probability of censoring weights, see Tian et al. (2014) or pseudo-observations, see Andersen et al. (2004).

The existing Bayesian research on RMST is limited to two approaches. Recently, Zhang and Yin (2023) proposed a Bayesian nonparametric analysis of RMST for right and interval-censored data by assigning a mixture of Dirichlet processes (MDP) prior to the distribution function. However, covariates adjustment is unavailable with this approach. Chen et al. (2023) overcomes this limitation, also considering a nonparametric dependent mixture model. In this paper, we extend the method of Andersen (2004) based on the pseudo-observations to the Bayesian framework, offering multivariable dRMST estimation without the need to model the survival function, providing an attractive alternative to existing Bayesian methods. This work follows a previous manuscript by Orsini et al. (2023) that is currently under review and discusses the Bayesian analysis of pseudo-observations with the Bayesian GMM to estimate hazard ratios.

# 2 Methods

Suppose that $\tilde{T}_i$ the time-to-event variable for the $i$-th subject, $Z_i$ a $p$-dimensional baseline covariate vector, $A_i$ the treatment allocation variable, and $C_i$ a right censoring random variable, independent of $\tilde{T}_i$, $Z_i$ and $A_i$. We observe $T_i = \tilde{T}_i \wedge C_i$ and $\Delta_i = I(\tilde{T}_i \leq C_i)$ the event indicator. For a pre-specified time point of interest $\tau$, the $\tau$-RMST is defined as:

$$\text{RMST}(\tau) = E(\tilde{T} \wedge \tau) = \int_0^\tau S(t)dt.$$

To adjust the RMST estimation on covariates, the following regression model can be considered:

$$\mu_i = E(\tilde{T}_i \wedge \tau | A, Z) = g^{-1}(\alpha + \delta A + \beta_1 Z_1 + \cdots + \beta_p Z_p)$$

where $g(\cdot)$ is a monotone differentiable link function and $\beta = (\alpha, \delta, \beta_1, \ldots, \beta_p)^T$ the vector of unknown parameters. With an identity link function, the regression coefficient, $\delta$, can be interpreted as the dRMST between the two arms.

We propose a Bayesian regression approach based on pseudo-observations to fit this model. Following Andersen et al. (2004), the $i$-th pseudo-observation is computed as

$$y_{\tau,i} = n \int_0^\tau \widehat{S}(t)dt - (n-1) \int_0^\tau \widehat{S}^{-i}(t)dt$$

with $n$ the sample size, $\widehat{S}(t)$ the Kaplan-Meier (KM) estimator of the survival probability, and $\widehat{S}^{-i}(t)$ the KM estimator excluding the $i$-th subject. Because of the unbiasedness of pseudo-observations conditional on covariates proved in Overgaard et al. (2017), we can replace the non-observed (due to censoring) $\tilde{T}_i \wedge \tau$ by $y_{\tau,i}$ in the regression model.

The Bayesian generalized method of moments (GMM) is used to estimate the posterior distribution $p(\beta|y_\tau) \propto \tilde{L}(\beta|y_\tau)p(\beta)$ where the pseudo-likelihood $\tilde{L}(\beta|y_\tau)$ is defined following

Yin (2009) as:

$$\tilde{L}(\beta|y_\tau) \propto \exp\{-\frac{1}{2}U_n^T(\beta)\Sigma_n^{-1}(\beta)U_n(\beta)\},$$

where

$$\Sigma_n(\beta) = \frac{1}{n^2}\sum_{i=1}^{n} u_i(\beta)u_i^T(\beta) - \frac{1}{n}U_n(\beta)U_n^T(\beta)$$

is a $(p+2)\times(p+2)$ matrix with $u_i(\beta) = \frac{\partial\mu_i}{\partial\beta}(y_{\tau,i} - \mu_i)$ and $U_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} u_i(\beta)$.

# 3 Simulation study

A simulation study of 2-arms randomized clinical trials with time-to-event outcomes was conducted to assess the performances of the Bayesian GMM based on pseudo-observations, to compare them with some of the RSMT estimators mentioned above (Andersen et al.(2004), Tian et al. (2014) and Zhang and Yin (2023)), and to evaluate the usefulness of covariates adjustment. Event times were simulated following a Weibull distribution, with scale and shape parameters chosen to mimic different patterns of treatment effect: PH (scenario 1), non-PH with early effect (scenarios 2 and 4), and delayed effect (scenarios 3 and 5). We also investigate the effects of covariate adjustment, drawn from uniform distribution (scenario 4) or normal and binomial distributions (scenario 5). Covariate effects were generated to be proportional. In all scenarios, a 30% censoring rate was considered, drawn from a uniform distribution with an administrative censoring at 8 years. Figure 1 displays the underlying survival curves for each scenario. The restriction time $\tau$ was set to 5 years, and 1000 replicates were generated for all scenarios. For the Bayesian approaches, noninformative priors $N(0, \sqrt{10}^2)$ were specified for all parameters of the Bayesian GMM regression model with pseudo-observations, and a mixture of Dirichlet processes prior was applied under an exponential base measure with Gamma $\Gamma(0.01, 0.01)$ mixing distribution for the method in Zhang and Yin (2023).

The main results are RMST estimation with covariates adjustment, corresponding to Scenario 4 with $n = 500$ (Table 1). The Bayesian GMM based on pseudo-observations gave similar results to the other methods. All the methods allowing for adjustment on covariates produced slightly more precise estimates. Similar results were observed in the other scenarios (data not shown).

# 4 Illustration on real data

For illustration, we analyzed the data from the Getug-AFU 15, a randomized, open-label, phase 3 trial comparing an androgen-deprivation therapy (ADT) alone ($n = 193$) or with docetaxel ($n = 192$) in non-castrate metastatic prostate cancer. The median follow-up time was 4.2 years. We focused on the Prostate-Specific Antigen (PSA) progression-free survival endpoint for which the PH hypothesis was rejected ($p = 0.00022$, Grambsch and Therneau). The survival curves for the two treatment groups show a diminution of the treatment effect

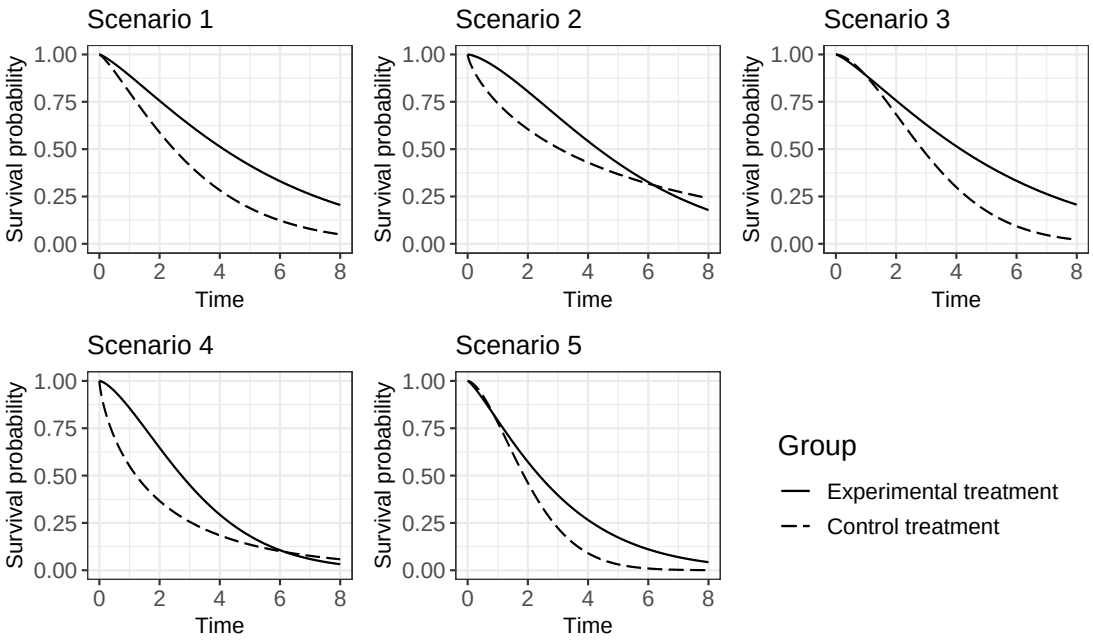



Figure 1: Theoretical survival curves for each simulated scenario.

| Methods | Bias | ASE[1] | RMSE[2] | 95% Coverage |
|---|---|---|---|---|
| **Frequentist** | | | | |
| Kaplan-Meier estimator | 0.0055 | 0.163 | 0.167 | 93.9 |
| Andersen et al. (2004)* | 0.0037 | 0.156 | 0.159 | 94.9 |
| Tian et al. (2014)* | 0.0032 | 0.155 | 0.159 | 94.9 |
| **Bayesian** | | | | |
| Zhang and Yin (2023) | 0.0054 | 0.162 | 0.167 | 93.9 |
| GMM* | 0.0060 | 0.156 | 0.158 | 94.7 |

[1] ASE: Average Standard Error, [2] RMSE: Root Mean Square Error

* Model adjusted on the prognostic variable $Z_1 \sim U([0, 2])$

Table 1: Performance of frequentist and Bayesian methods for the estimation of the difference of 5-RMST between 2 arms in Scenario 4.

over time (Figure 2). The difference in 5-RMST was estimated around 0.58 year for all methods (Figure 3), meaning that receiving docetaxel in addition to ADT increases the life time without PSA progression during the next 5 years by 0.58 a year, compared to receiving ADT alone. Estimating unadjusted 5-RMST with the Bayesian GMM approach based on pseudo-observations gave similar results compared to the existing methods. After adjustment on four prognostic variables (gleason score, European Cooperative Oncology Group performance status, concentration of alkaline phosphatase, and presence of bone metastases), in complete-case, an increase in precision was observed.

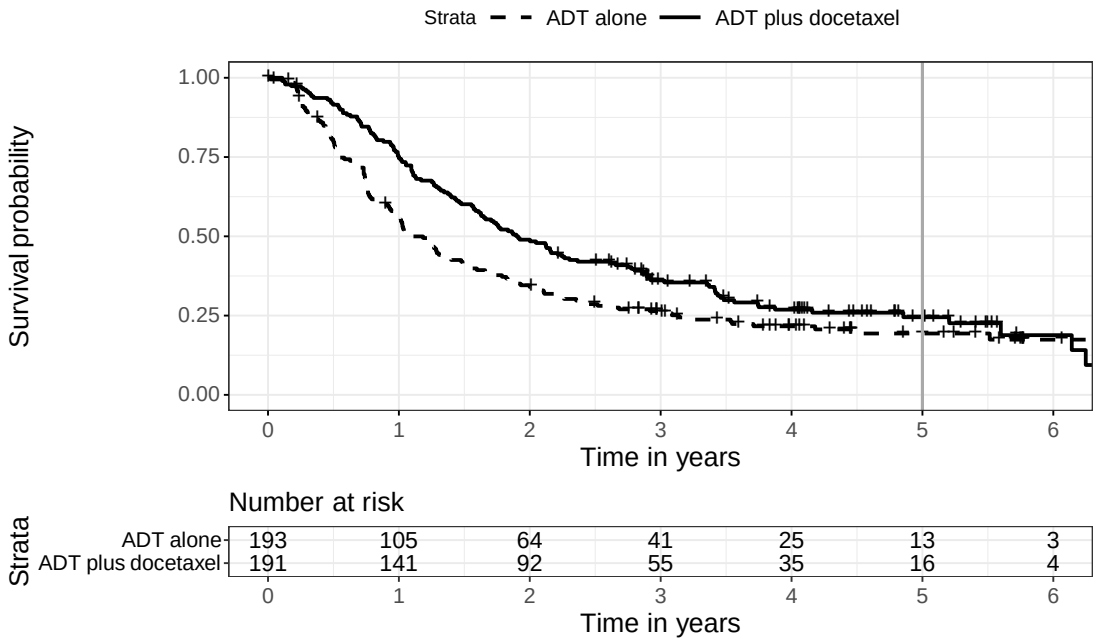



Figure 2: Kaplan-Meier curves for the PSA progression-free survival from the Getug-AFU 15 trial. The vertical grey line represents the restricted time $\tau = 5$ year.

# 5 Discussion

This paper introduces a novel Bayesian approach for analyzing RMST, based on pseudo-observations and the Bayesian generalized methods of moments. The first advantage is to eliminate the need to specify and integrate the survival function by fitting a regression model on the RMST directly, making this method more straightforward compared to existing Bayesian approaches. The second advantage is to provide a direct estimation of RMST adjusted on covariates in the Bayesian framework. Caution must be taken regarding the potential misspecification of the model, especially with continuous covariates, since the functional form of the relationship between the covariate and outcome must be specified. The third advantage is to provide the probability of the dRMST being above any desired cut-off, given from the posterior distribution.

Whatever the methods used for RMST analysis, the critical point is the pre-specification of the time point of interest $\tau$. To avoid this arbitrary choice, Ambrogli et al. (2022) developed RMST curve estimations based on pseudo-observations. Further research will extend the Bayesian GMM based on pseudo-observations to jointly analyze RMST at multiple time points. This can be achieved by considering a vector of pseudo-observations.

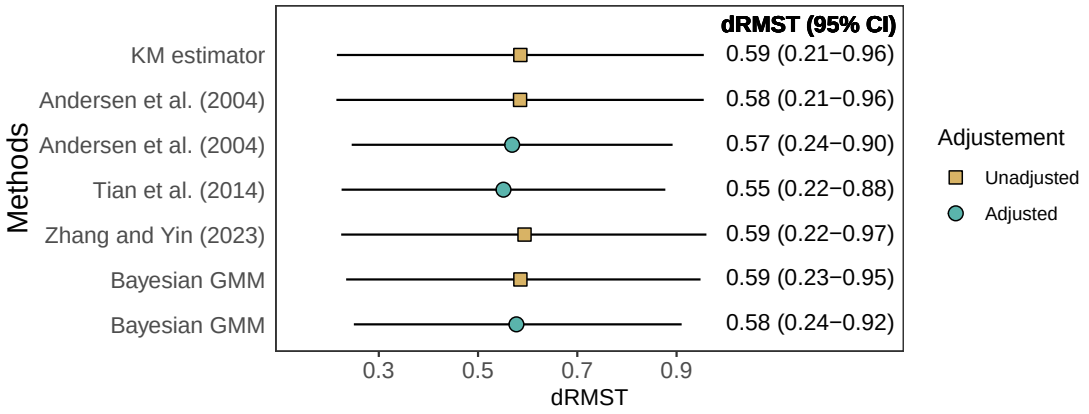



Figure 3: Estimation of the difference of 5-RMST between the ADT plus docetaxel and ADT alone groups for the PSA progression-free survival from the Getug-AFU 15 trial, the horizontal lines represent the 95% confidence or credibility intervals.

# Bibliographie


Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004), Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations. *Lifetime Data Analysis*, 10(4), pp. 335-350.

Ambrogi, F., Iacobelli, S., and Andersen, P. K. (2022). Analyzing differences between restricted mean survival time curves using pseudo-values. *BMC Medical Research Methodology*, 22(1), pp. 71.

Chen, R., Basu, S., and Shi, Q. (2023). Restricted Mean Survival Time Estimation Using Bayesian Nonparametric Dependent Mixture Models. *arXiv* (arXiv:2305.14639).

Karrison, T. (1987). Restricted Mean Life With Adjustment for Covariates. *Journal of the American Statistical Association*, 82(400), pp. 1169-1176.

Karrison, T., and Kocherginsky, M. (2018), Restricted mean survival time: Does covariate adjustment improve precision in randomized clinical trials? *Clinical Trials*, 15(2), pp. 178-188.

Orsini, L., Brard, C., Lesaffre, E., Guosheng, Y., Dejardin, D., and Le Teuff, G. (2023) Bayesian generalized method of moments applied to pseudo-observations in survival analysis. *Submitted for review*

Overgaard, M., Parner, E. T., and Pedersen, J. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, 45(5), 1988–2015.

Tian, L., Zhao, L., and Wei, L. J. (2014), Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15(2), pp. 222-233.

Yin, G. (2009), Bayesian generalized method of moments. *Bayesian Analysis*, 4(2), pp. 191-

208.

Zhang, C., and Yin, G. (2023), Bayesian nonparametric analysis of restricted mean survival time. *Biometrics*, 79(2), pp. 1383-1396.

Zucker, D. M. (1998), Restricted Mean Life with Covariates: Modification and Extension of a Useful Survival Analysis Method.*Journal of the American Statistical Association*, 93(442), pp. 702–709.

# Pseudo-Observations and Super Learner for the Estimation of the Restricted Mean Survival Time

Ariane Cwiling [1] & Vittorio Perduca [1] & Olivier Bouaziz [1]

[1] *Université Paris Cité, MAP5, UMR 8145, France*

*ariane.cwiling@u-paris.fr* *vittorio.perduca@parisdescartes.fr*
*olivier.bouaziz@parisdescartes.fr*

**Résumé.** Il peut être utile pour les cliniciens de prédire le délai avant l'apparition d'un événement tel qu'une rechute, l'apparition d'un cancer ou le décès d'un patient. En présence de données censurées à droite, il est naturel de considérer plutôt un délai restreint en raison des problèmes d'estimation de la queue de distribution. Le problème de prédiction est alors équivalent à l'estimation du temps de survie moyen restreint. À cette fin, nous proposons un nouveau modèle de régression flexible et facile à utiliser, basé sur les pseudo-observations et le super learner. Pour prouver la validité théorique de cette méthode, nous présentons une nouvelle définition des pseudo-observations.

**Mots-clés.** Analyse de survie, RMST, pseudo-observations, super learner, prédiction.

**Abstract.** It can be relevant for clinicians to have access to a prediction of the time to an event such as a relapse, a cancer occurrence, or the death of a patient. When predicting the time to event based on right-censored data, it is natural to rather consider a restricted time because of tail estimation issues. The prediction task is then equivalent to the estimation of the restricted mean survival time (RMST). To that aim, we propose a new flexible and easy to use regression model based on pseudo-observations and super learning. To prove the theoretical validity of this method, we present a new definition of the pseudo-observations.

**Keywords.** Survival analysis, RMST, pseudo-observations, super learner, prediction.

## 1 Introduction

Survival analysis on right-censored data typically involves estimating hazard and survival functions. However, there is a growing interest in directly predicting the time to event, particularly in medical contexts such as predicting relapse, cancer occurrence, or patient death. Due to tail estimation challenges, predicting up to a relevant fixed time horizon, or equivalently estimating the restricted mean survival time (RMST), is preferred. The RMST is a clinically meaningful quantity that has gained attention for its simplicity and interpretability. In particular, pseudo-observations, introduced by Andersen et al. (2004), have enabled the application of a large range of prediction models by transforming incomplete observed times into data that can be handled as uncensored. This approach facilitates the use of various

prediction models adapted to uncensored data, from generalized linear models (Andersen et al., 2004) to neural networks (Zhao, 2021). The best prediction model in a defined library can be selected with cross-validation (Van Der Laan and Dudoit, 2003). A simple improvement to this method is not to select the best model but the best combination of them, called super learning (Van Der Laan et al., 2007). Previous research has explored super learning for analyzing censored data, such as the survival super learner proposed by Golmakani and Polley (2020), albeit with limitations regarding the proportional hazard assumption. Another example is the super learner with Inverse Probability Censoring Weight (IPCW) loss (Keles et al., 2004 ; Devaux et al., 2022), which requires consistent censoring estimation. In this study, we investigate the predictive ability of super learning applied to pseudo-observations for right-censored data, which remains unexplored. To adapt the convergence result of the super learner from Van Der Laan et al. (2007), we introduce a new type of pseudo-observations, called *split* pseudo-observations, primarily for theoretical purposes, as they exhibit similar practical performance to classic pseudo-observations.

# 2 Pseudo-observations

In the context of right-censored data, we denote by $T^*$ the variable of interest, $C$ the censoring time, $T = T^* \wedge C$ the observed variable and $\delta = \mathbb{1}\{T^* \leq C\}$ the censoring indicator. An observation is represented by the vector $O = (T, \delta, Z)$ where $Z \in \mathbb{R}^d$ is a covariate vector. We note $S(t \mid Z) = \mathbb{P}(T^* > t \mid Z)$ the survival function of $T^*$ conditionally on the covariates $Z$. Let $\tau_H = \inf\{t > 0 : \mathbb{P}(T > t \mid Z) = 0 \text{ a.s.}\}$. The RMST is defined for a fixed time horizon $\tau \leq \tau_H$, conditionally on the covariates, as

$$\mathbb{E}[T^* \wedge \tau \mid Z] = \int_0^\tau S(t \mid Z)dt.$$

Given this definition, the RMST can be estimated for instance by integrating an estimator of the survival function between 0 and $\tau$, or by regressing the restricted event times on covariates. In the second case, censoring must be taken into account since the times $T^*$ are not observed for all individuals. This can be achieved by using pseudo-observations. Consider censored observations $D_n = \{O_i = (T_i, \delta_i, Z_i)\}$, $i = 1, \ldots, n$. Classical pseudo-observations are computed in the following way, for a given $\tau < \infty$,

$$\Gamma_i := n \int_0^\tau \hat{S}(t)dt - (n-1) \int_0^\tau \hat{S}^{-i}(t)dt, \ i = 1, \ldots, n, \tag{1}$$

where $\hat{S}$ is the Kaplan-Meier estimator of the survival function computed on all data and $\hat{S}^{-i}$ is the same estimator computed on all data but the $i$-th. The interest of pseudo-observations for regression purposes lies in the following result by Jacobsen and Martinussen (2016),

$$\mathbb{E}[\Gamma_i \mid Z_i] = \mathbb{E}[T^* \wedge \tau \mid Z_i] + \mathbb{E}[\xi_n \mid Z_i], \tag{2}$$

where $\xi_n = o_{\mathbb{P}}(1)$. This result is valid under the following independent censoring assumption.

**Assumption 1** (Independent censoring)**.** *The censoring variable $C$ and the pair of variables $(T^*, Z)$ are independent.*

Pseudo-observations are, by construction, correlated with each other, which makes it difficult to study their theoretical properties. To deal with this issue, we propose a new type of pseudo-observations, called *split pseudo-observations*. The idea is to split the data in two subsets $D_{n_1}$ and $D_{n_2}$ of size $n_1$ and $n_2 = n - n_1$, respectively. The former is used to compute the Kaplan-Meier estimator and the latter for the pseudo-observations. We then define a new type of pseudo-observations as follows :

$$\Gamma_i(D_{n_1}) = \Gamma_{O_i}(D_{n_1}) := (n_1 + 1) \int_0^\tau \hat{S}_{D_{n_1},O_i}(t)dt - n_1 \int_0^\tau \hat{S}_{D_{n_1}}(t)dt, \ i = 1, \dots, n_2, \tag{3}$$

where $O_i$ represents an observation from $D_{n_2}$, $\hat{S}_{D_{n_1}}$ is the Kaplan-Meier estimator of the survival function computed on the $n_1$ data points in $D_{n_1}$ and $\hat{S}_{D_{n_1},O_i}$ is the same estimator computed on the $n_1 + 1$ data points obtained by adding $O_i$ to the sample $D_{n_1}$. The main advantage of this construction is that the new pseudo-observations constructed for all observations in $D_{n_2}$ are independent conditionally on $D_{n_1}$. A result similar to Equation (2) can then be easily derived for those split pseudo-observations. Under Assumption 1 :

$$\mathbb{E}[\Gamma_i(D_{n_1}) \mid Z_i, D_{n_1}] = \mathbb{E}[T^* \wedge \tau \mid Z_i] + \mathbb{E}[\xi_{n_1} \mid Z_i, D_{n_1}], \tag{4}$$

where $\xi_{n_1} = o_{\mathbb{P}}(1)$. In Section 3 we establish a convergence result for the super learner coupled with split pseudo-observations. Simulation results are presented in Section 4. In practice, we observe that split and traditional pseudo-observations are very similar, and the choice between them has minimal impact on the prediction quality. Therefore, split pseudo-observations are mostly used for theoretical purposes while traditional pseudo-observations are mostly used in applications.

# 3 Super Learning

The super learner algorithm is based on cross-validation. During this process, a new dataset is constructed where each observation is paired with a set of predictions from all candidate learners. The most effective algorithm, determined by comparing predictions to observations using a specific loss function, is designated as the discrete super learner. On the other hand, the continuous super learner (simply termed as the "super learner" in what follows) derives the optimal combination of models using a user-chosen algorithm to assign weights to all candidate learners. Theorem 1 in Dudoit and Van Der Laan (2005) shows that the discrete super learner performs asymptotically as well as, or better, than any of the candidate learners, when using a quadratic loss - the mean squared error (MSE). The extension of this result to the continuous super learner is immediate, for it only involves considering the minimum cross-validated risk predictor as based on a parametric regression, as outlined by Van Der Laan et al. (2007). Our aim is to adapt those results to our method, using the MSE as the loss function. This loss allows to compare a wide range of models without making any specific model assumption. Besides, it is well known that, under this loss, the best prediction model is the conditional expectation of the variable of interest, which, in our case, results in estimating the RMST. In order to take into account right-censored data, a first approach is to use an IPCW loss (Keles et al., 2004 ; Devaux et al.,

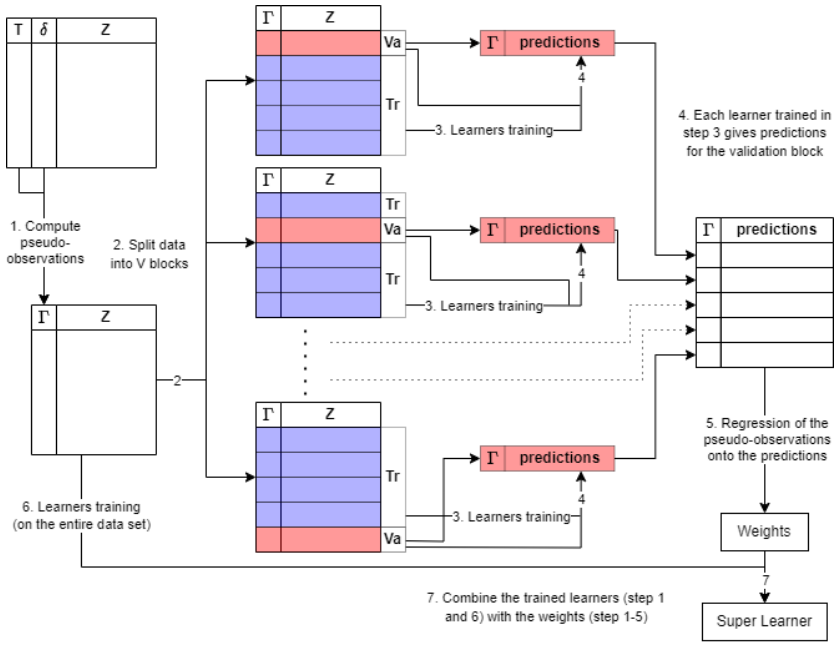



FIGURE 1 – Diagram of the super learner based on standard pseudo-observations for right-censored data, see Equation (1). Pseudo-observations are computed once and for all at the beginning.

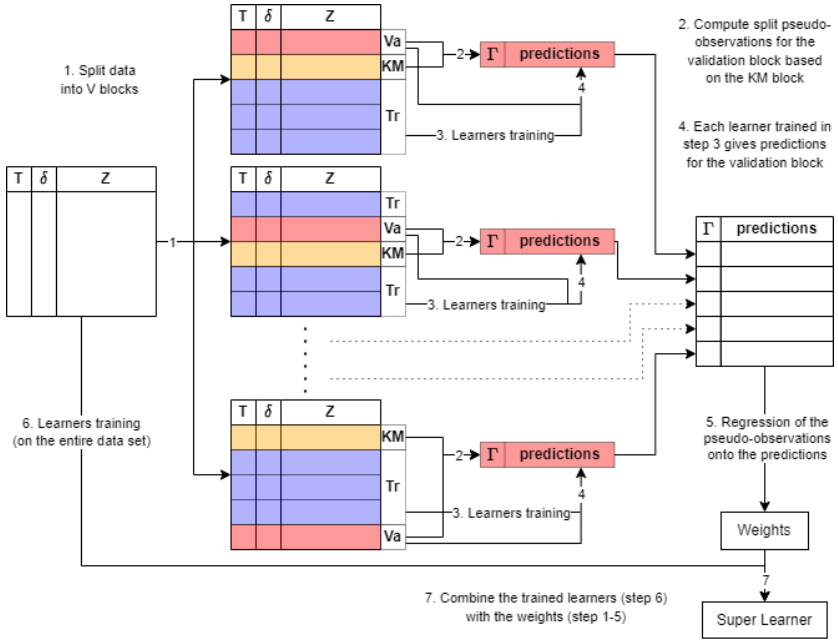



FIGURE 2 – Diagram of the super learner based on split pseudo-observations for right-censored data, see Equation (3). Pseudo-observations are computed during the cross-validation step, for each validation block, based on an additional subset of the data (KM).

2022). Another idea, proposed by Golmakani and Polley (2020), is to minimize the negative log partial likelihood. Our approach consists in feeding pseudo-observations directly into the super learner, using a standard quadratic loss applied on those pseudo-observations. This idea is motivated by the asymptotic result in Equation (2). Our algorithm is therefore identical to the super learner described in Van Der Laan et al. (2007), with an additional first step for computing pseudo-observations. A diagram illustrating the method is provided in Figure 1. However, the dependence structure of the pseudo-observations makes it difficult to provide theoretical results with this method. This is why we also introduce a different algorithm, described in Figure 2, for which we can prove the classical optimality result for the super learner by taking advantage of the conditional independence structure of the split pseudo-observations. The algorithm is slightly different in that case, as it requires using a subset of the data to compute pseudo-observations in the validation set. In practice, we will show in Section 4 that applying the super learner to standard or split pseudo-observations does not make a significant difference in terms of predictive performance. Thus, the theoretical results below are provided with regard to the algorithm based on split pseudo-observations (Figure 2), yet we recommend the use of standard pseudo-observations in practice (Figure 1) for ease of use, lower computational cost and allocation of more data to learners training.

Formally, consider triplets $\{(T_i^*, C_i, Z_i)\}$, $i = 1, \dots, n$, $T_i^* \in \mathbb{R}^+$ the time to event, $C_i^* \in \mathbb{R}^+$ the censoring time, $Z_i \in \mathbb{R}^d$ the covariates, of joint law $P(t) = \mathbb{P}(T^* \leq t, C \leq c, Z \leq z)$. We only observe a transformation of these triplets defined as the data set $D_n = \{O_i = (T_i, \delta_i, Z_i) = (T_i^* \wedge C_i, \mathbb{1}\{T_i^* \leq C_i\}, Z_i)\}$, $i = 1, \dots, n$, $T_i \in \mathbb{R}^+$, $\delta_i \in \{0, 1\}$, $Z_i \in \mathbb{R}^d$. We consider $K_n$ candidate estimators for the RMST $\widehat{\psi}_k$, $k = 1, \dots, K_n$, among which we wish to select the best in terms of risk, using cross-validation. Combining split pseudo-observations with cross-validation imposes to divide the data in three subsets instead of two. Observations are divided according to an independent random vector $B_n = (B_n(i) : i = 1, \dots, n) \in \{0, 1, 2\}^n$ into a first training set $\{O_i : i, B_n(i) = 0\}$ of size $n_0 = n - \lfloor np_{1,n} \rfloor - \lfloor np_{2,n} \rfloor$ for $p_{1,n}, p_{2,n}, p_{1,n} + p_{2,n} \in (0, 1)$, a second training set $\{O_i : i, B_n(i) = 1\}$ of size $n_1 = \lfloor np_{1,n} \rfloor$ and a validation set $\{O_i : i, B_n(i) = 2\}$ of size $n_2 = \lfloor np_{2,n} \rfloor$. The first training set is used to construct the candidate estimators of the RMST. The second training set is used to compute the Kaplan-Meier estimator which in turn is used for the computation of the pseudo-observations. We refer to this set as the Kaplan-Meier (KM) set. Pseudo-observations are computed for the data in the validation set. We denote $P_{B_n}^0$, $P_{B_n}^1$ and $P_{B_n}^2$ the empirical distributions of the three subsets. Several cross-validation schemes, i.e. distributions for $B_n$, exist. We focus on $V$-fold cross-validation, where data are divided into $V$ subsets, or folds, of approximately same size. One by one, each fold serves as a validation set while the remaining folds constitute the training sets. The associated distribution of $B_n$ assigns a mass of $1/V$ to each of the $V$ binary vectors. In this context, Equation (4) can be rewritten as

$$\mathbb{E}[\Gamma_O(P_{B_n}^1) \mid Z, P_{B_n}^1, B_n] = \mathbb{E}[T^* \wedge \tau \mid Z] + \mathbb{E}[\xi_{n_1} \mid Z, P_{B_n}^1, B_n], \tag{5}$$

where $\Gamma_O(P_{B_n}^1)$ is the split pseudo-observation constructed for the observation $O$ based on the distribution $P_{B_n}^1$ of the KM set. Consider the quadratic loss function for pseudo-observations

$$L^{\text{po}} : (\psi, P_{B_n}^1, O) \mapsto \big(\Gamma_O(P_{B_n}^1) - \psi(Z)\big)^2, \text{ for a parameter } \psi, P_{B_n}^1 \sim P^{\otimes n_1}, O \sim P.$$

The quantity of interest is the *risk* of the parameter $\psi$ for the distributions $P$ and $P^1_{B_n}$,

$$\Theta^{\text{po}}(\psi, P^1_{B_n}, P) = \int L^{\text{po}}(\psi, P^1_{B_n}, o) dP(o).$$

Given this definition, the (unknown) risk minimizer is defined as

$$\psi^*_1 = \psi^*_1(P^1_{B_n}, P) = \arg\min_{\psi \in \Psi} \Theta^{\text{po}}(\psi, P^1_{B_n}, P) = \arg\min_{\psi \in \Psi} \int L^{\text{po}}(\psi, P^1_{B_n}, o) dP(o),$$

and characterizes the optimal risk

$$\theta^*_1 = \Theta^{\text{po}}(\psi^*_1, P^1_{B_n}, P) = \min_{\psi \in \Psi} \Theta^{\text{po}}(\psi, P^1_{B_n}, P) = \min_{\psi \in \Psi} \int L^{\text{po}}(\psi, P^1_{B_n}, o) dP(o).$$

The cross-validated risk estimator for the $k$-th candidate learner is defined as

$$\begin{aligned}
\widehat{\theta}^{\text{po}}_n(k) &= \mathbb{E}_{B_n} \Theta^{\text{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, P^2_{B_n}\big) \\
&= \mathbb{E}_{B_n} \int L^{\text{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, o\big) dP^2_{B_n}(o) \\
&= \mathbb{E}_{B_n} \frac{1}{n_2} \sum_{i: B_n(i)=2} L^{\text{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, O_i\big).
\end{aligned}$$

We wish to select the learner that minimizes this risk. This cross-validated selector is denoted

$$\widehat{k}^{\text{po}} = \arg\min_{k \in \{1, \ldots, K_n\}} \widehat{\theta}^{\text{po}}_n(k).$$

Optimality results are based on the comparison between the cross-validation selector and the selector which for each given data set makes the best choice, knowing the true data distribution (Van Der Laan and Dudoit, 2003). This cross-validated oracle selector minimizes the cross-validated conditional risk

$$\widetilde{\theta}^{\text{po}}_n(k) = \mathbb{E}_{B_n} \Theta^{\text{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, P\big) = \mathbb{E}_{B_n} \int L^{\text{po}}\big(\widehat{\psi}_k(P^0_{B_n}), P^1_{B_n}, o\big) dP(o),$$

and is denoted

$$\widetilde{k}^{\text{po}} = \arg\min_{k \in \{1, \ldots, K_n\}} \widetilde{\theta}^{\text{po}}_n(k).$$

We compared the risk differences $\widetilde{\theta}^{\text{po}}_n(\widehat{k}^{\text{po}}) - \theta^*_1$ and $\widetilde{\theta}^{\text{po}}_n(\widetilde{k}^{\text{po}}) - \theta^*_1$ to demonstrate optimality in a manner similar to Theorem 1 in Dudoit and Van Der Laan (2005).

**Theorem 1.** *Let $O_1, \ldots, O_n$ be a random sample from a data generating distribution $P$. Each $O_i = (T_i, \delta_i, Z_i)$ consists of a univariate outcome $T_i \in \mathbb{R}^+$, a binary censoring indicator $\delta_i \in \{0, 1\}$ and a covariate vector $Z_i \in \mathbb{R}^d$. Let $\{\widehat{\psi}_k : k = 1, \ldots, K_n\}$ denote a sequence of $K_n$ candidate estimators for the RMST, $\mathbb{E}[T^* \wedge \tau \mid Z]$. If we consider the quadratic loss function $L^{po}(\psi, P^1_{B_n}, O) = \big(\Gamma_O(P^1_{B_n}) - \psi(Z)\big)^2$, then the risk minimizer $\psi^*_1(Z) = \mathbb{E}\big[\Gamma_O(P^1_{B_n}) \mid Z, P^1_{B_n}, B_n\big]$ is asymptotically equivalent to the RMST (see Equation (5)). Suppose that*

$$|\Gamma_O(P^1_{B_n})| \leq M < \infty \text{ and } \sup_{Z, \psi \in \Psi} |\psi(Z)| \leq M < \infty \text{ almost surely,}$$

*where the supremum is taken over a support of the distribution of $Z$. Suppose that Assumption 1 holds.*

***Finite sample result.*** *Let $M_1 = 8M^2$, $M_2 = 16M^2$ and $c(M,\gamma) = 2(1+\gamma)^2(M_1/3 + M_2/\gamma)$. For all $\gamma > 0$,*

$$0 \leq \mathbb{E}[\widetilde{\theta}_n^{po}(\widehat{k}^{po}) - \theta_1^*] \leq (1+2\gamma)\mathbb{E}[\widetilde{\theta}_n^{po}(\widetilde{k}^{po}) - \theta_1^*] + 2c(M,\gamma)\frac{1+\log(K_n)}{n_2}.$$

***Asymptotic result.***

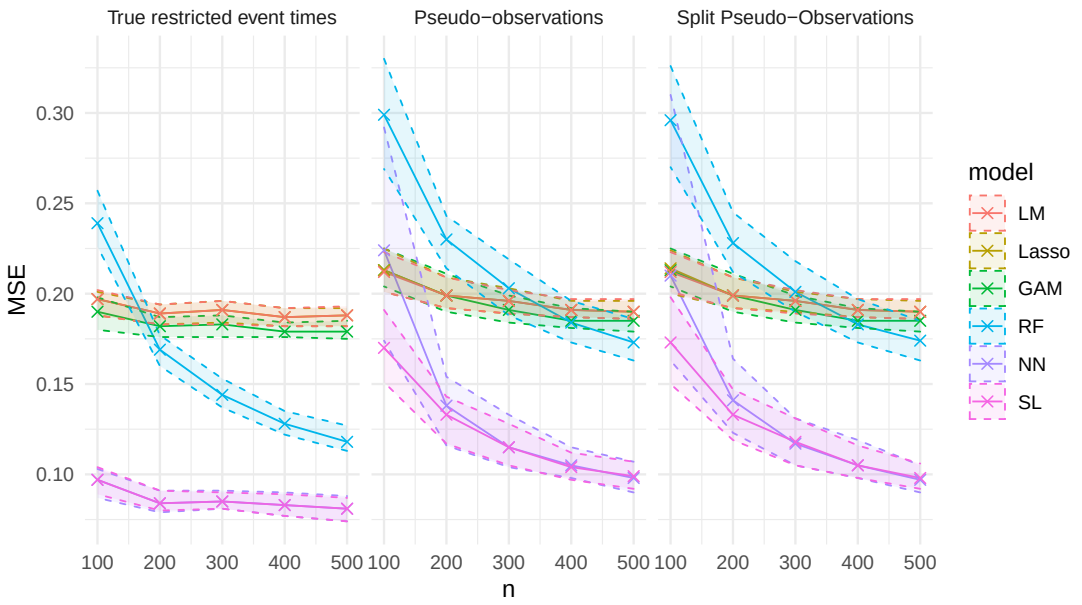
$$\textit{If} \quad \frac{\log(K_n)}{n_2(\widetilde{\theta}_n^{po}(\widetilde{k}^{po}) - \theta_1^*)} \xrightarrow[n\to\infty]{\mathbb{P}} 0, \quad \textit{then} \quad \frac{\widetilde{\theta}_n^{po}(\widehat{k}^{po}) - \theta_1^*}{\widetilde{\theta}_n^{po}(\widetilde{k}^{po}) - \theta_1^*} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

# 4 Simulations

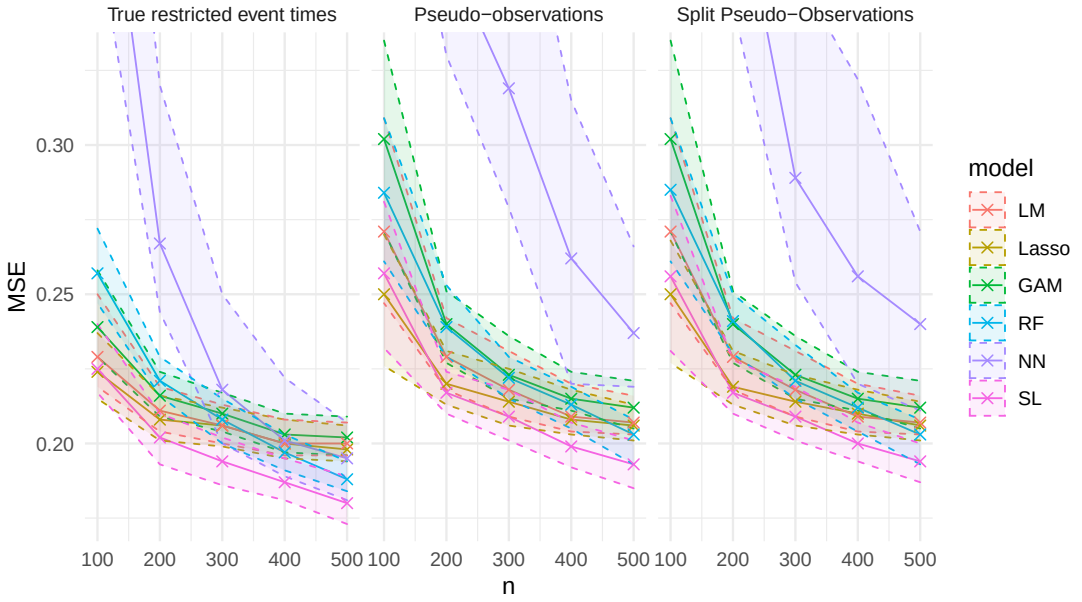
In this section, we present some simulation results that validate our approach. Data are simulated according to two different settings, 1 and 2, corresponding respectively to schemes B and C in Cwiling (2023). Both rely on the Cox model, with different levels of complexity.

Combining standard pseudo-observations with super learning simply implies computing pseudo-observations on the whole data set and feeding them into the super learner as outcomes (see Figure 1). Standard prediction models can be directly applied to the pseudo-observations, thus we chose as candidate learners the linear model (LM), the Lasso, the Generalized Additive Model (GAM), the Random Forest (RF) and the Neural Network (NN). On the other hand, split pseudo-observations require a different methodology as illustrated in Figure 2. In particular, the learners applied to the training set need to handle right-censored observations. In order to be able to compare this methodology with the previous one, we therefore chose to consider candidate learners based on pseudo-observations, combined with LM, Lasso, GAM, RF, and NN.

We simulated data of increasing size $n \in \{100, 200, 300, 400, 500\}$ on which we applied both types of pseudo-observations based super learners. For comparison, we added the classic super learner trained on the true restricted event times. All super learners were trained using 6-folds cross-validation. The MSEs of the super learner and of every candidate learners were computed on an independent test set of size 1000. We repeated the process 80 times for each data size $n$, simulating new data sets each time. The performances of the candidate learners were evaluated from their last training on the whole data set. The results displayed in Figure 3 show similar performances for both types of pseudo-observations, regardless of the simulation scheme, indicating no significant difference in practice. Figure 3 also illustrates the asymptotic result from Theorem 1, stating that the super learner applied to split pseudo-observations performs asymptotically as well as or better than any of the candidate learners. This also seems to be the case with standard pseudo-observations, even though we did not provide any theoretical result in that case. Thus, in practice, we recommend to compute standard pseudo-observations once and for all and feed them into the super learner. This reduces the complexity of the method and allows to allocate more data to the training of candidate learners.

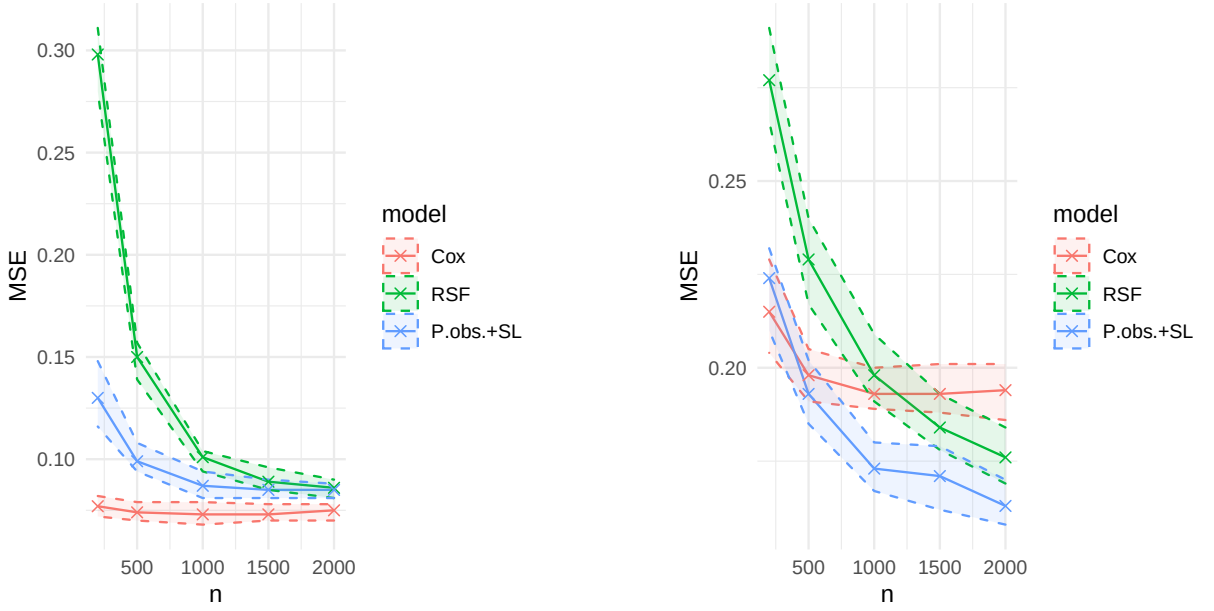



(a) Simulation scheme 1




(b) Simulation scheme 2

FIGURE 3 – Application of the super learner on standard and split pseudo-observations and on true restricted event times, for data sets of size $n$. The MSEs of the super learner and of every candidate learners were computed on a test set of size 1000. The process was repeated 80 times for each data size $n$, with new data sets each time. Values of the median (cross), first and third quantiles (dashed lines) are reported.




(a) Simulation scheme 1

(b) Simulation scheme 2

FIGURE 4 – Prediction of the restricted time to event with the Cox model (no interaction between covariates is included), the RSF and pseudo-observations combined with super learning. The algorithms were trained on a data set of size $n$. The MSEs were computed on a test set of size 1000. The process was repeated 80 times for each data size $n$, with new data sets each time. Values of the median (cross), first and third quantiles (dashed lines) are reported.

As a consequence, in the next simulations, we only selected the method with traditional pseudo-observations and compared its predictive performance with other survival methods. Our method was implemented with the same candidate learners and same number of folds as previously, and compared to two widely used survival methods, namely the Cox model (including no interaction between covariates) and the Random Survival Forests (RSF). These methods are used to estimate the survival function, from which the estimation of the RMST is derived by integrating the survival curve between 0 and $\tau$. We simulated data of increasing size $n \in \{200, 500, 1000, 1500, 2000\}$ on which we applied the three methods. The MSEs were computed on an independent test set of size 1000. We repeated the process 80 times for each data size $n$, simulating new data sets each time. The results are displayed in Figure 4. The first simulation scheme consists in simulating time to events according to a Cox model without interactions between covariates. Hence, as expected, Figure 4a shows a better performance for the Cox model compared to the other methods. However, the RSF and our method show similar performances for large sample sizes, with our approach providing a better predictive performance than the RSF on the overall. In the second simulation scheme, interactions between covariates are included in the Cox model used to simulate the data. The results are presented in Figure 4b for the Cox model (without interactions), the RSF and our method. It is seen that our method outperforms the Cox model for $n \geq 500$ and the RSF for all sample sizes.

# 5 Conclusion

This study introduces a novel approach for predicting the restricted time to event from right-censored data by combining pseudo-observations with the super learner. Numerical experiments demonstrated that our method performs at least as well as the best candidate learner and competes favorably with classical survival methods like the Cox model or RSF. The independent censoring assumption and its computation cost are the main limitations of the method. On the other hand, the combination of pseudo-observations with the super learner is a straightforward method and its good results on simulated and real data motivates its use in practice. To take into account the dependence structure of the pseudo-observations, we proposed the conditionally independent split pseudo-observations which allowed us to extend the convergence result of the super learner to right-censored data. The potential behind the split pseudo-observations goes beyond its application to the super learner : it can be used in various fields to address theoretical issues related to their dependence structure.

# Bibliographie


Andersen, P. K., Hansen, M. G., and Klein, J. (2004), Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations, *Lifetime Data Analysis*, 10(4), pp. 335-350.

Cwiling, A., Perduca, V. and Bouaziz, O. (2023), A Comprehensive Framework for Evaluating Time to Event Predictions using the Restricted Mean Survival Time.

Devaux, A., Genuer, R., Peres, K. and Proust-Lima, C. (2022), Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach, *BMC Medical Research Methodology*, 22(1), pp. 188.

Dudoit, S. and Van Der Laan, M. J. (2005), Asymptotics of cross-validated risk estimation in estimator selection and performance assessment, *Statistical Methodology*, 2(2), pp. 131-154.

Golmakani, M. K. and Polley, E. C. (2020), Super Learner for Survival Data Prediction, *The International Journal of Biostatistics*, 16(2), pp. 20190065.

Jacobsen, M. and Martinussen, T. (2016), A Note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-observations, *Scandinavian Journal of Statistics*, 43(3), pp. 845-862.

Keles, S., Van Der Laan, M. J. and Dudoit, S. (2004), Asymptotically optimal model selection method with right censored outcomes, *Bernoulli*, 10(6).

Van Der Laan, M. J. and Dudoit, S. (2003), Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated adaptive Epsilon-Net Estimator : Finite Sample Oracle Inequalities and Examples, *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Van Der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007), Super Learner, *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Zhao, L. (2021), Deep neural networks for predicting restricted mean survival times, *Bioinformatics*, 36(24), pp. 5672-5677.

# Données fonctionnelles 2

# Directional regularity: Achieving faster rates of convergence in multivariate functional data

Omar Kassi*, Sunny Wang*

February 12, 2024

**Abstract**

We introduce a new notion of regularity, called *directional regularity*, which is relevant for a wide range of applications involving multivariate functional data. We show that for anisotropic functional data, faster rates of convergence can be obtained by adapting to its directional regularity through a change of basis. An algorithm is constructed for the estimation and identification of the directional regularity for a large class of stochastic processes, made possible due to the unique replication nature of functional data. Accompanying non-asymptotic theoretical guarantees are provided. A novel simulation algorithm, which is of independent interest, is designed to evaluate the numerical accuracy of our directional regularity algorithm. Simulation results demonstrate the good finite sample properties of our estimator, whose implementation is freely available in the form of a **R** package.

**Abstract**

Nous présentons une nouvelle notion de régularité, appelée *régularité directionnelle*, qui s'avère pertinente pour une vaste gamme d'applications impliquant des données fonctionnelles multivariées. Nous démontrons qu'en adaptant la base aux données fonctionnelles anisotropes, des taux de convergence optimal peuvent être atteints en exploitant leur régularité directionnelle. Nous développons un algorithme pour estimer et identifier cette régularité directionnelle pour une large classe de processus stochastiques, ce qui est rendu possible par la nature de réplication unique des données fonctionnelles. Nous fournissons également des garanties théoriques non asymptotiques. De plus, nous concevons un nouvel algorithme de simulation, d'intérêt indépendant, pour évaluer la précision numérique de notre méthode d'estimation. Les résultats de simulation illustrent les bonnes propriétés en un ensemble de données fini de notre estimateur, dont l'implémentation est disponible sous forme d'un package **R**.

*Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France; omar.kassi@ensai.fr

*Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France; sunny.wang@ensai.fr

Cette étude a bénéficié d'une aide de l'Etat gérée par l'ANR dans le cadre du projet France 2030 EUR DIGISPORT (ANR-18-EURE-0022)

# 1 Introduction

## 1.1 Motivation

The prevalence of multivariate functional data analysis (fda) is burgeoning in many fields, ranging from healthcare to environmental science. For an exposition or survey of fda, see Ramsay and Silverman (2005), Hsing et al. (2016), Wang et al. (2016), among other excellent sources. It is well known by now that the rates of convergence for estimating many diverse quantities in fda depend on the underlying smoothness of the process. Some references include Chagny and Roche (2016), Cai and Yuan (2011), Cai and Yuan (2012), Golovkine et al. (2023), Wang Guang Wei et al. (2023). Despite the progress of adaptive estimation in fda, much of the current adaptive estimation literature in the functional data landscape falls within the univariate setting, since the multivariate framework brings along its own set of challenges. As a simple motivating example, let us go back to classical, one curve multivariate non-parametric regression setup. In this setting, pairs $(X_i, Y_i), i = 1, \ldots, n$ are observed under the model

$$Y_i = f(X_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $f : [0,1]^d \to \mathbb{R}$ and the $X_i$'s are independent design points uniformly distributed on the hypercube $[0,1]^d$ and the $\epsilon_i$'s are uncorrelated, centered random variables. Assuming that the regression function $f$ belongs to the anisotropic Hölder class, it is well known that under suitable assumptions on the noise structure (Hoffmann and Lepski (2002)), the minimax rate of estimation is $n^{-\gamma/(2\gamma+1)}$, where $\gamma$ is also known as the effective smoothness, given by

$$\frac{1}{\gamma} = \sum_{i=1}^{d} \frac{1}{\gamma_i},$$

where $\gamma_i$ is the regularity along dimension $i$. In the non-parametric estimation literature, the "maximising" smoothness $\gamma_i$ is assumed to intrinsically exist in the directions of the canonical bases. In the functional data setting where one is for instance interested in the smoothing of surfaces such as images, Lemma 1 shows that this is an unreasonably restrictive assumption.

Lemma 1 basically states that in general, there is only one direction (among infinitely many) in which the "maximising" regularity lies, and one is often paying the price of the worst regularity in each dimension by working only in the canonical bases. Going back to our motivating example of non-parametric regression, if the maximising regularities are indeed not in the direction of the canonical bases, the effective smoothness that one often obtains is then instead given by

$$\gamma = \frac{\min_{i=1,\ldots,d} \gamma_i}{d}$$

thereby only obtaining estimation rates of which correspond to the isotropic case. To further illustrate the implications of Lemma 1, we now provide a concrete motivating example of an anisotropic stochastic process, restricting ourselves to the class of processes that satisfy condition (2). For any point $\mathbf{t} \in \mathcal{T}$, we denote $(t_1, t_2)$ to be the coordinates of $\mathbf{t}$ in the $(\mathbf{u_1}, \mathbf{u_2})$ basis, where $\mathbf{u_i}, i = 1, 2$ are unit vectors in the unit circle $\mathbb{S}$ that spans $\mathbb{R}^2$. Let

$H_1 < H_2, H_i \in (0,1), i = 1,2$, and $B_1, B_2$ be two independent fractional brownian motion with Hurst indexes $H_1$ and $H_2$ respectively. Define the following process, which is the sum of two fractional brownian motions (fBms):

$$X(\mathbf{t}) = B_1(t_1) + B_2(t_2), \qquad \forall \mathbf{t} \in \mathcal{T}. \tag{1}$$

It is well known that for any $\Delta > 0$, we have

$$\mathbb{E}\left[\{B_i(\mathbf{t} - \Delta/2) - B_i(\mathbf{t} + \Delta/2)\}^2\right] = \Delta^{2H_i}, \qquad \forall t \in \mathbb{R}_+.$$

The independence of $B_1$ and $B_2$ implies that

$$\mathbb{E}\left(\{X(\mathbf{t} - \Delta/2\mathbf{u_i}) - X(\mathbf{t} + \Delta/2\mathbf{u_i})\}^2\right) = \Delta^{2H_i},$$

so the sum of the regularities when working in the $(\mathbf{u_1}, \mathbf{u_2})$ basis is $H_1 + H_2 > 2H_1$, where the right-hand side of the inequality corresponds to the isotropic case. Several other such examples exist, and we can in fact quantify that these class of processes are sufficiently large. In order to profit from the inherent anisotropy of processes such as (1), we introduce a new notion of regularity in the functional data setting which takes into account the underlying anisotropy. We call this concept *directional regularity*, which is defined by the map $\mathbf{v} \mapsto H_{\mathbf{v}}$, inspired by the fact that they characterise the "anisotropic smoothness" of the process. A formal definition can be found in Definition 1. Our goal in this paper to construct an algorithm which locates the directional vector $\mathbf{v}^* = \arg\max_{\mathbf{v} \in \mathbb{S}} H_{\mathbf{v}}$; this is equivalent to finding the angle $\alpha \in [0, 2\pi]$ (the range $[0, 2\pi]$ can be reduced to $[0, \pi)$ since $H_{-\mathbf{v}} = H_{\mathbf{v}}$, for any $\mathbf{v} \in \mathbb{S}$) between $\mathbf{v}^*$ and the first canonical basis $\mathbf{e}_1$ (In the bivariate case, the second maximising direction then simply lies in a $\pi/2$ reflection of $\alpha$). Detecting the angle $\alpha$ allows one to perform a change-of-basis from the canonical ones where the data is observed, to the ones which provide the directions of the maximising regularities. The anisotropy of the process can then be exploited, thus obtaining faster convergence rates. An illustration can be seen in Figure 1.

# 2 Methodology

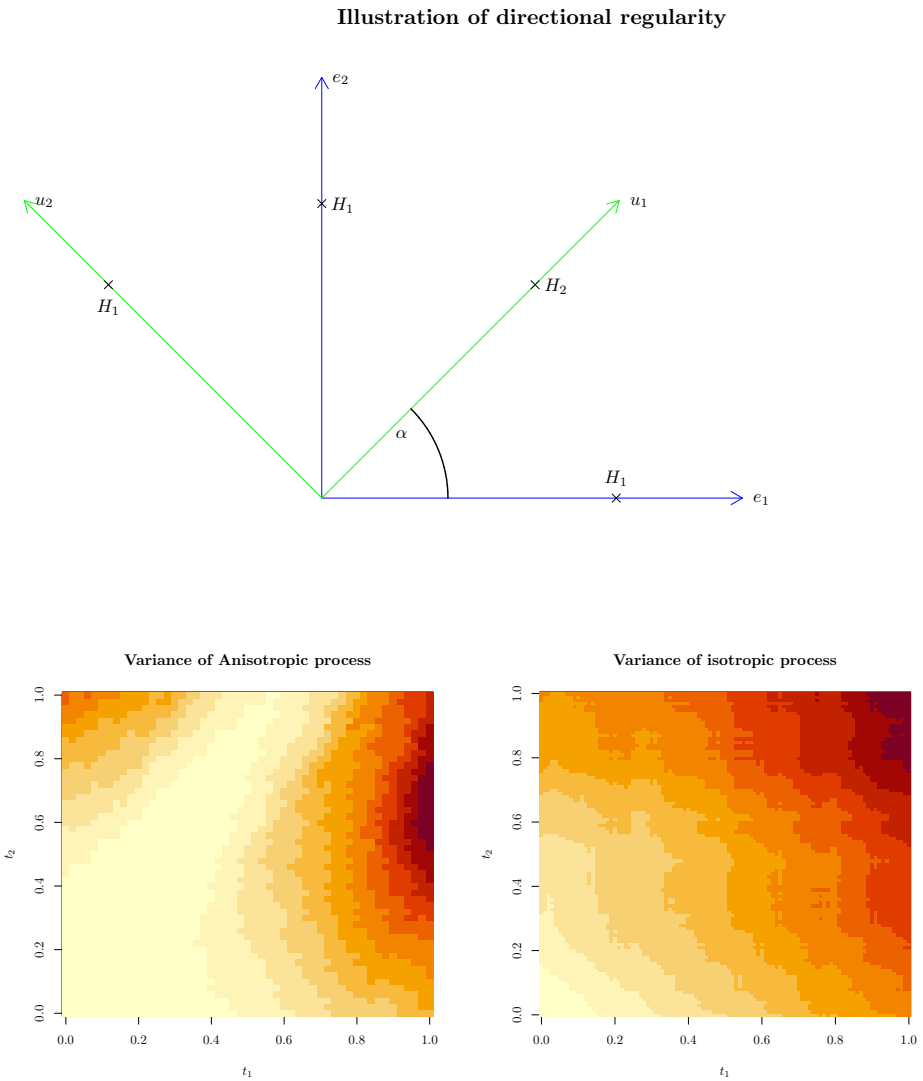
## 2.1 Data Setting

We suppose that the data observations $(Y^{(j)}(\mathbf{t}_m), \mathbf{t}_m)$ are generated from the following model

$$Y^{(j)}(\mathbf{t}_m) = X^{(j)}(\mathbf{t}_m) + \epsilon^{(j)}(\mathbf{t}_m), \qquad 1 \leq j \leq N, \leq m \leq M_0, \ \mathbf{t}_m \in \mathcal{T},$$

such that the errors are independent, centered random variables with constant variance $\sigma^2$. Although our approach can be generalised to the random design case where the time points $\mathbf{t}_m$ are drawn from some distribution and vary between individuals, together with heteroscedastic noise structures, we focus on the common design and homoscedastic case in our exposition below for the sake of clarity.

Figure 1: Isotropic rates are obtained when working in the canonical basis when there is anisotropy (here $H_1 < H_2$). The second plot shows the variance of an anisotropic fractional brownian sheet, while the third plot displays the variance of an isotropic process.



## 2.2 Problem formulation

The notion of local regularity in quadratic mean was considered by Golovkine et al. (2022) in the case of one dimension, and extended by Kassi et al. (2023) in the multivariate case. Our notion of directional regularity stands on the shoulders of these previous results, which we will formally introduce in this subsection. We recall that the domain $\mathcal{T}$ is a bivariate subset of $\mathbb{R}^2$, and we denote with $\mathbb{S}$ the unit circle of $\mathbb{R}^2$.

**Definition 1.** *Let $u \in \mathbb{S}$, $X$ a non-differentiable stochastic process and $H_u : \mathcal{T} \to (0,1)$. We say that the process $X$ has a local regularity $H_u$ in a point $\mathbf{t} \in \mathcal{T}$ along the direction $\mathbf{u}$ if a bounded function $L_u : \mathcal{T} \to \mathbb{R}_+$ exist such that :*

$$\theta_{\mathbf{u}}(\mathbf{t}, \Delta) := \mathbb{E}\left[\left\{X\left(\mathbf{t} - \frac{\Delta}{2}u\right) - X\left(\mathbf{t} - \frac{\Delta}{2}u\right)\right\}^2\right] = L_u(\mathbf{t})\Delta^{2H_u(\mathbf{t})} + G(\mathbf{t}, \Delta), \tag{2}$$

*where $G(\mathbf{t}, \Delta) \underset{\Delta \to 0}{=} o\left(\Delta^{2H_u(\mathbf{t})}\right)$.*

The key difference in Definition 1 with the local regularity introduced in Kassi et al. (2023) is that in 1, for each fixed direction $\mathbf{u}$, a local regularity is associated with it and found in the first dominating term of $\theta_{\mathbf{u}}(\mathbf{t}, \Delta)$. On the other hand, the framework of Kassi et al. (2023) supposes that at most two regularities exist and is found in the first two dominating terms of $\theta_{\mathbf{e}_1}(\mathbf{t}, \Delta)$ and $\theta_{\mathbf{e}_2}(\mathbf{t}, \Delta)$. Thus the concept of directional regularity allows one to restrict the study of regularities up to the first order term of $\theta_{\mathbf{u}}$ instead of the second order term, a much easier estimation problem.

If the function $H_u$ does not depend on the direction $u$, we say that $X$ is an isotropic process, otherwise we call it an anisotropic process. An example of an anisotropic process with prescribed directional regularity is provided in (1). In this paper we consider $G(\mathbf{t}, \Delta) = \Delta^{2H_{\mathbf{u}}(\mathbf{t}) + \beta(\mathbf{t})}$ for some positive function $\beta > 0$. As mentioned in the previous sections, the following Lemma illustrates the importance of directional regularity.

**Lemma 1.** *Assume that there exists basis vectors $(\mathbf{u_1}, \mathbf{u_2}) \in \mathbb{S}$ that spans $\mathbb{R}^2$ such that $H_{\mathbf{u_1}} < H_{\mathbf{u_2}}$. Moreover, suppose that the functions $H_{\mathbf{u_1}}, H_{\mathbf{u_2}}, L_{\mathbf{u_1}}, L_{\mathbf{u_2}}$ and $\beta$ are continuously differentiable. For any $\mathbf{v} \in \mathbb{S}$, we have the following dichotomy:*

- *If $\mathbf{v} \neq \pm\mathbf{u_2}$, then the regularity along $\mathbf{v}$ is $H_{\mathbf{u_1}}$.*
- *Otherwise, the local regularity along $\mathbf{v}$ is $H_{\mathbf{u_2}}$.*

The previous lemma shows that the map $\mathbf{v} \mapsto H_{\mathbf{v}}$ can only take at most two possible values. Furthermore, the maximisation problem $\arg\max_{\mathbf{v} \in \mathbb{S}} H_{\mathbf{v}}$ admit two solutions $\mathbf{u_2}$ and -$\mathbf{u_2}$, which means that we can restrict ourselves to the upper plane of the unit circle. The problem is then equivalent to $\arg\max_{\alpha \in [0,\pi)} H_{\mathbf{u}(\alpha)}$, where $\mathbf{u}(\alpha) = \cos(\alpha)\mathbf{e}_1 + \sin(\alpha)\mathbf{e}_2$.
In what follows, we only consider the case of constant regularity. (i.e the function $H_{\mathbf{u}}$ does not vary along the domain $\mathcal{T}$). The general case can be obtained by considering a local, pointwise study.

## 2.3 Estimating equations

Herein we assume that $X$ is an anisotropic process. Let $(\mathbf{e_1}, \mathbf{e_2})$ be the canonical basis of $\mathbb{R}^2$, and $(\mathbf{u_1}, \mathbf{u_2})$ be orthonormal basis vectors such that $\mathbf{u_1}$ or $\mathbf{u_2}$ maximise the directional regularity. Let $\alpha \in [0, \pi)$ be the angle between the two basis vectors $\mathbf{u_1}$ and $\mathbf{e_1}$ so that $\langle \mathbf{e_1}, \mathbf{u_1} \rangle = \cos(\alpha)$. The following proposition provides the estimating equation of the angle $\alpha$.

**Proposition 1.** *Suppose that for $i = 1, 2$, $\mathbf{e}_i \notin \arg\max_{\mathbf{v} \in \mathbb{S}} H_{\mathbf{v}}$. Let $H_i$ denote the regularity along $\mathbf{u_i}$ for $i = 1, 2$. Then we have*

$$|g(\alpha)| = |g(\alpha, \Delta)| = \left( \frac{\theta_{\mathbf{e_2}}(\mathbf{t}, \Delta)}{\theta_{\mathbf{e_1}}(\mathbf{t}, \Delta)} \right)^{\frac{1}{2\underline{H}}} + O\left( \Delta^{\beta \wedge 2H_{\mathbf{e}} \wedge 1} \right),$$

*where $g = \tan \mathbf{1}\{H_1 < H_2\} + \cot \mathbf{1}\{H_1 > H_2\}$, and $\underline{H} = \min\{H_1, H_2\}$.*

That is, the angles can be computed, up to a reflection, by taking the ratios of mean-squared variations given by (2) along the directions of the canonical basis. Due to the unique replication nature of function data, this quantity is easily estimable. A natural plug-in estimator is given by

$$\widehat{\theta}_{e_i}(\mathbf{t}, \Delta) = \frac{1}{N} \sum_{j=1}^{N} \left\{ \widetilde{X}^{(j)}\left(\mathbf{t} - (\Delta/2)\mathbf{e_i}\right) - \widetilde{X}^{(j)}\left(\mathbf{t} + (\Delta/2)\mathbf{e_i}\right) \right\}^2, \qquad i = 1, 2,$$

where $\widetilde{X}^{(j)}$ denotes some observable approximation of $X^{(j)}$. One can choose between a variety of methods, ranging from simplest ones such as interpolation, to slightly more complex non-parametric smoothers, as long as $R_2(\mathfrak{m}) \leq \mathcal{L}\mathfrak{m}^{-\nu}, \forall \mathfrak{m} \geq 1$ is satisfied, where $R_p(\mathfrak{m}) = \sup_{\mathbf{t} \in \mathcal{T}} \mathbb{E}[|\widetilde{X}_j(\mathbf{t}) - X_j(\mathbf{t})|^p]$, a mild condition satisfied by most non-parametric smoothers (e.g Fan and Guerre (2016)). The regularity $H_{\mathbf{v}}(\mathbf{t})$ can be estimated as follows:

$$\widehat{\underline{H}} = \begin{cases} \min_{i=1,2} \frac{\log(\widehat{\theta}_{\mathbf{e_i}}(\mathbf{t}, 2\Delta)) - \log(\widehat{\theta}_{\mathbf{e_i}}(\mathbf{t}, \Delta))}{2\log(2)} & \text{if} \quad \widehat{\theta}_{\mathbf{e_i}}(\mathbf{t}, 2\Delta), \widehat{\theta}_{\mathbf{e_i}}(\mathbf{t}, \Delta) > 0, \\ 1 \quad \text{otherwise.} \end{cases}$$

The regularity estimator above is a slight adaptation of Kassi et al. (2023) by taking the minimum over the index of the basis vectors. By collecting the two estimators above, we thus obtain the plug-in estimator

$$g^{-1}\left|\widehat{g(\alpha)}\right| = g^{-1}\left( \frac{\widehat{\theta}_{\mathbf{e_2}}(\mathbf{t}, \Delta)}{\widehat{\theta}_{\mathbf{e_1}}(\mathbf{t}, \Delta)} \right)^{\frac{1}{2\widehat{\underline{H}}}}. \tag{3}$$

Since (3) holds for any $\mathbf{t} \in \mathcal{T}$ in our setting, and $\alpha$ does not vary along the domain, we suggest to average each estimate over the grid of points $\mathbf{t}$ to obtain more stable estimates. In this case, the only parameter that needs to be chosen in (3) is $\Delta$. A principled selection of $\Delta$ is provided in this work.

## 2.4 Identification issues

Proposition 1 reveals that estimation of the angle $\alpha$ possesses two identification problems, arising from the phenomenon that the dominating term arising from the ratio of mean squared variations differ depending on the intrinsic directional regularities. An algorithm is provided in this research work for the identification process but left our due to space constraints; here we simply point out that is works very well in practice.

# 3 Theoretical Properties

Under mild assumptions, the following results hold true.

**Theorem 1.** *Two positive constant $C_1$ and $C_2$ exist such that for any $u \in (0,1)$ we have*

$$\mathbb{P}\left(|\widehat{g(\alpha,\Delta)} - g(\alpha,\Delta)| \geq u\right) \leq C_1 \exp\left(-C_2 N u^2 \frac{\Delta^{2H}}{\log^2 \Delta}\right),$$

*where $g$ is defined in Proposition 1.*

**Corollary 1.** *We have the following rates of convergence for $\widehat{\alpha}$:*

$$|\widehat{\alpha}(\Delta) - \alpha| = O_{\mathbb{P}}\left(\max\left\{\frac{1}{\min\{\sqrt{N}, \mathfrak{m}^H\}}, \frac{|\log \Delta|}{\sqrt{N}\Delta^H}\right\}\right).$$

# 4 Numerical Properties

Our simulation study was conducted using a new simulator which enables one to simulate a bivariate anisotropic fractional brownian sheet. To the best of our knowledge, we do not know any existing methods that allows one to easily simulate anisotropic processes. Due to space constraints, we will not describe the simulator in this abstract, and simply present some early simulation results.

## 4.1 Parameter settings and error measures

Observations $(Y^{(i)}(\mathbf{t}_m^{(i)}), \mathbf{t}_m), 1 \leq i \leq N, 1 \leq m \leq M_0$ were simulated using our novel algorithm for the sum of two fBms $f_1(B_1, B_2) = B_1 + B_2$. A total of 48 different parameter configurations were explored, consisting of all possible combinations of the following parameter sets: number of curves $N \in \{100, 200\}$, number of points along each curve $M_0 \in \{26^2, 51^2\}$, noise level $\sigma \in \{0, 0.01, 0.05, 0.1\}$, and angles $\alpha \in \{\pi/3, \pi/5, 5\pi/6\}$. We fixed the regularities to be $H_1 = 0.8$ and $H_2 = 0.5$; the robustness of identification when the dominating term changes is taken into account by considering the experiment $\alpha = 5\pi/6 > \pi/2$, which is equivalent to taking a case when $H_1 < H_2$. Since we observed that the estimation of $\alpha$ were fairly robust to the choice of spacing parameter $\Delta$ as long as it is large enough, we decided

to select $\Delta = M_0^{-1/4}(1 + \Delta_c)$, where $\Delta_c = 0.25$. Since this choice consistently performs well in all our configurations, we are fairly comfortable recommending it as a "universal" choice for the purposes of $\widehat{\alpha}$. On the other hand, the grid of spacings $\mathbf{\Delta}$ taken in the identification process were $\mathbf{\Delta} = \{M^{-1/4}, \Delta_1, \ldots, \Delta_{k-1}, 0.4\}$, an evenly spaced grid such that the cardinality $\#\mathbf{\Delta} = 15$, which seems to be sufficiently large. Once the final $\widehat{\alpha}$ were obtained by running our estimation and identification algorithms on the simulated data sets, we then computed the absolute error as a risk measure for each experiment:

$$\mathcal{R}_\alpha = |\widehat{\alpha} - \alpha|.$$

## 4.2 Empirical Results

Results can be seen in Figures 2 and 3 in the form of boxplots. We can see at first glance that our estimator performs relatively well, and that the maximum risk stays below 0.15, where higher risk values are observed for $\alpha = 5\pi/6$. For smaller values of $\alpha$, for example when $\alpha = \pi/6$, we can see that the risk values largely stays below 0.05, except for the setups with very large noise. Perhaps what might seem surprising is that when the number of observed points along each curve are small, (e.g $M = 26$), the risk decreases as the noise level increases for some values of $\alpha$. This is probably due to the fact that the estimation error resulting from noise is being dominated by that arising from being sparsely observed. This aligns with the results that we observe for larger sample sizes, since we see that even in the presence of significant noise, the risk is still lower any of those experiments associated with the lower number of sampling points.

Figure 2: Boxplots for $N = 100$ curves

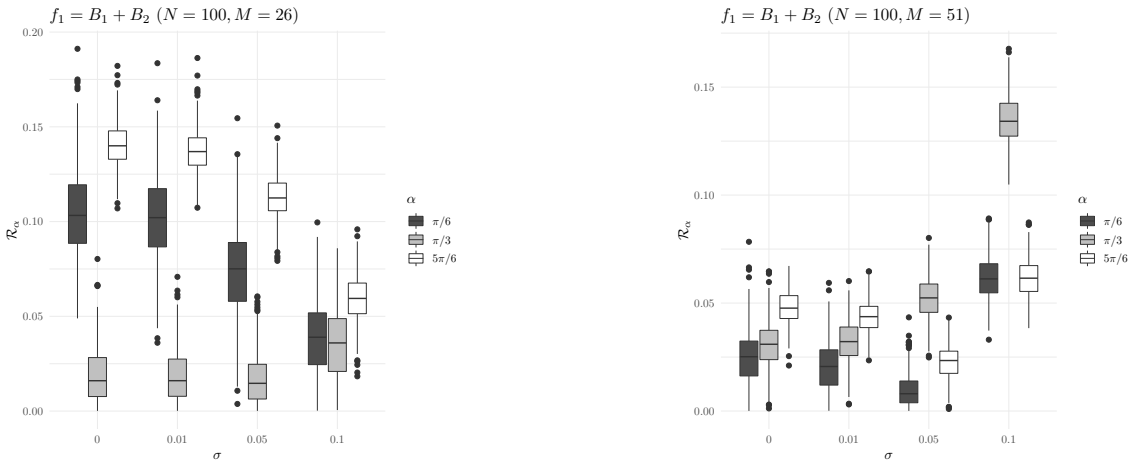
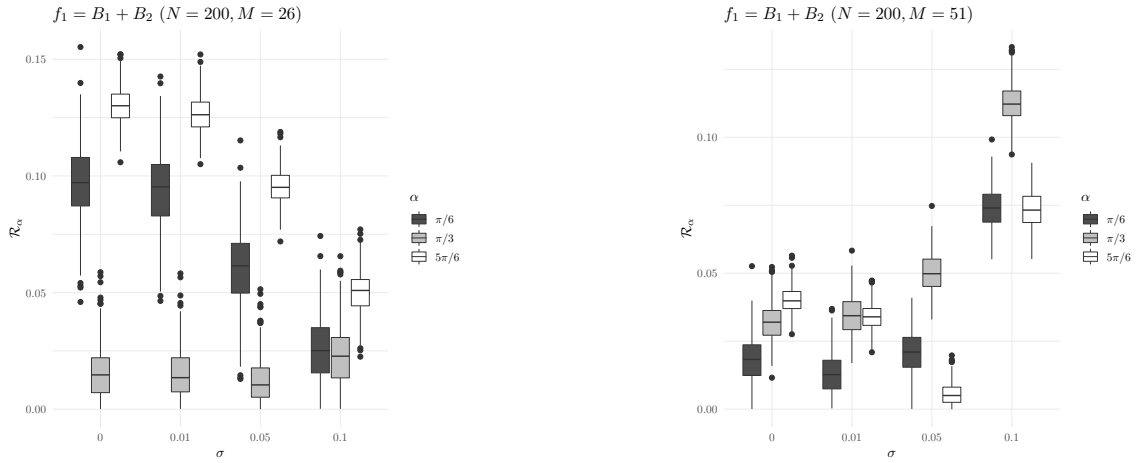


Figure 3: Boxplots for $N = 200$ curves




# References


Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Statist.*, 39(5):2330–2355.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216.

Chagny, G. and Roche, A. (2016). Adaptive estimation in the functional nonparametric regression model. *Journal of Multivariate Analysis*, 146:105–118. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

Fan, Y. and Guerre, E. (2016). Multivariate local polynomial estimators: Uniform boundary properties and asymptotic linear representation. In *Essays in Honor of Aman Ullah*, volume 36, pages 489–537. Emerald Group Publishing Limited.

Golovkine, S., Klutchnikoff, N., and Patilea, V. (2022). Learning the smoothness of noisy curves with application to online curve estimation. *Electron. J. Stat.*, 16(1):1485–1560.

Golovkine, S., Klutchnikoff, N., and Patilea, V. (2023). Adaptive estimation of irregular mean and covariance functions. arxiv 2108.06507v2.

Hoffmann, M. and Lepski, O. (2002). Random rates in anisotropic regression. (With discussion). *Ann. Stat.*, 30(2):325–396.

Hsing, T., Brown, T., and Thelen, B. (2016). Local intrinsic stationarity and its inference. *The Annals of Statistics*, 44(5):2058 – 2088.

Kassi, O., Klutchnikoff, N., and Patilea, V. (2023). Learning the regularity of multivariate functional data.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295.

Wang Guang Wei, S., Patilea, V., and Klutchnikoff, N. (2023). Adaptive functional principal components analysis. arxiv 2306.16091.

# Régression additive sous variables imparfaites

Germain Van Bever [1] & Jeong Min Jeon [2]

[1] *Université libre de Bruxelles et Université de Namur, Belgique, germain.van.bever@ulb.be*
[2] *Seoul National University, Corée du Sud, jeongmin.jeon.stat@gmail.com*

**Résumé.** Dans cet exposé, nous présentons un modèle additif dans lequel la variable réponse prend ses valeurs dans un espace de Hilbert. Les prédicteurs sont multivariés. Toutes les variables peuvent possiblement être imparfaitement mesurées. Le modèle permet de considérer des variables réponses Euclidiennes, fonctionnelles ou même à valeur dans des espaces de densité. L'ajout de variables imparfaites permet de couvrir le cas de variables fonctionnelles à valeurs dans une variété Riemannienne, le cas où seul un échantillon aléatoire d'une densité inconnue est disponible, ou encore le cas où les régresseurs sont les scores obtenus par analyse en composantes principales ou singulières dans un espace de Hilbert. L'estimation des fonctions de régression se fait via la méthode de smooth backfitting (Mammen *et al.*, 1990). Nous étudions le comportement non-asymptotique et asymptotique de ces estimateurs. Plusieurs applications illustrent les méthodes introduites.

**Mots-clés.** Analyse des données fonctionnelles, Erreurs de mesure, Espaces de Hilbert, Régression additive

**Abstract.** In this talk, we study an additive model where the response variable is Hilbert-space-valued and predictors are multivariate Euclidean, and both are possibly imperfectly observed. Considering Hilbert-space-valued responses allows to cover Euclidean, compositional, functional and density-valued variables. By treating imperfect responses, we can cover functional variables taking values in a Riemannian manifold and the case where only a random sample from a density-valued response is available. Dealing with imperfect predictors allows us to cover various principal component and singular component scores obtained from Hilbert-space-valued variables. For the estimation of the additive model having such variables, we use the smooth backfitting method originated by Mammen et al. (1999). We provide full non-asymptotic and asymptotic properties of our regression estimator and present its wide applications via several simulation studies and real data applications..

**Keywords.** Additive regression, Functional data analysis, Hilbert valued data, measurement errors

# 1 Modèle additif hilbertien

Le modèle de régression additive "classique", pour des régresseurs $X_j \in \mathbb{R}$, $j = 1, \ldots, d$, et une variable réponse $Y \in \mathbb{R}$, prend la forme

$$Y = f_0 + \sum_{j=1}^{d} f_j(X_j) + \epsilon,$$

où $f_0 \in \mathbb{R}$ est une constante et les fonctions $f_j$, $j = 1, \ldots, d$ sont les fonctions-composantes inconnues à estimer. Ce modèle permet une balance entre un modèle paramétrique, peu flexible, et un modèle totalement nonparamétrique du type $Y = f(X_1, \ldots, X_d) + \epsilon$, dans lequel les taux de convergence de l'estimateur de $f$ souffrent du "curse of dimensionality".

Soit $\mathbb{H}$ un espace de Hilbert séparable. Soit également $\oplus$, $\odot$, $\mathbf{0}$, $\langle \cdot, \cdot \rangle$ et $\|\cdot\|$, respectivement, l'addition vectorielle, la multiplication scalaire, le vecteur nul, le produit scalaire et la norme sur $\mathbb{H}$. Notons que $\oplus$, $\odot$, $\mathbf{0}$, $\langle \cdot, \cdot \rangle$ et $\| \cdot \|$ pour $\mathbb{H} = \mathbb{R}^k$ correspondent à $+$, $\times$, $(0, \ldots, 0) \in \mathbb{R}^k$, le produit scalaire et la norme $\ell_2$, respectivement. Des exemples d'espaces $\mathbb{H}$ non-euclidiens sont disponibles dans les exemples de la section 3.

Dans ce papier, nous considérons le modèle additif Hilbertien multivarié

$$\mathbf{Y} = \mathbf{f}_0 \oplus \bigoplus_{j=1}^{d} \mathbf{f}_j(\xi_j) \oplus \epsilon, \tag{1}$$

où $\mathbf{Y} \in \mathbb{H}$ est une variable réponse satisfaisant $\mathbb{E}(\|\mathbf{Y}\|^2) < \infty$, $\xi_j = (\xi_{j1}, \ldots, \xi_{jL_j}) \in \mathbb{R}^{L_j}$ avec $L_j \in \mathbb{N}$ sont des prédicteurs multivariés, $\epsilon \in \mathbb{H}$ est un terme d'erreur satisfaisant $\mathbb{E}(\|\epsilon\|^2) < \infty$ et $\mathbb{E}(\epsilon | \xi_1, \cdots, \xi_d) = \mathbf{0}$, $\mathbf{f}_0 \in \mathbb{H}$ est une constante inconnue et $\mathbf{f}_j : \mathbb{R}^{L_j} \to \mathbb{H}$ sont des fonctions-composantes inconnues. Ici, l'espérance conditionnelle $\mathbb{E}(\epsilon | \xi_1, \cdots, \xi_d)$ est définie via une intégrale de Bochner.

Nous supposons également que les régresseurs $\xi_{jl}$, pour $1 \leq j \leq d$ et $1 \leq l \leq L_j$, peuvent provenir de différentes sources. Par exemple, $\xi_{jl}$ peuvent être des prédicteurs scalaires ou des scores obtenus par une analyse en composantes principales ou singulières. Ces prédicteurs scalaires peuvent, parfois, souffrir d'erreurs de mesures (Delaigle, 2008). Les scores principaux sont inobservables en général, la vrai structure de covariance étant inconnue. Pour ces raisons, nous supposons que seule une approximation imparfaite $\tilde{\xi}_{jl} \in \mathbb{R}$ de $\xi_{jl}$ est disponible. Pour les mêmes raisons, nous supposons observer une approximation $\tilde{\mathbf{Y}}$ de $\mathbf{Y}$. Ceci couvre, entre autres, des réponses fonctionnelles reconstruites sur base d'évaluations temporelles, etc.

# 2 Estimation

L'estimation des fonctions-composantes $\mathbf{f}_j$, $j = 1, \ldots, d$ sur un domaine compact $D_j \subset \mathbb{R}^{L_j}$ se fait via la méthode de smooth backfitting (Mammen, 1999). Soit $p$ la densité de $\xi = (\xi_1, \ldots, \xi_d)$ et soit $p_0^D = \int_D p(\mathbf{x}) d\mathbf{x} > 0$, où $D = \prod_{j=1}^{d} D_j$. Soit $p^D(\mathbf{x}) = p(\mathbf{x})/p_0^D$ pour $\mathbf{x} = (x_1, \ldots, x_d) \in D$. Soit la densité marginale $p_j^D(x_j) = \int_{D_{-j}} p^D(\mathbf{x}) d\mathbf{x}_{-j}$ et $p_{jk}^D(x_j, x_k) =$

$\int_{D_{-jk}} p^D(\mathbf{x})d\mathbf{x}_{-jk}$, où $D_{-j} = \prod_{m\neq j} D_m$, $D_{-jk} = \prod_{m\neq j,k} D_m$, et $\mathbf{x}_{-j}$ et $\mathbf{x}_{-jk}$ denotent, respectivement, les $(d-1)$- et $(d-2)$-vecteurs obtenus en omettant $x_j$ et $(x_j, x_k)$ in $\mathbf{x}$.

Remarquons que les $\mathbf{f}_j$ ne sont pas identifiables dans le modèle (1) puisque $\bigoplus_{j=0}^d \mathbf{f}_j = \bigoplus_{j=0}^d (\mathbf{f}_j \oplus \mathbf{c}_j)$ pour toutes constantes $\mathbf{c}_j \in \mathbb{H}$ telles que $\bigoplus_{j=0}^d \mathbf{c}_j = \mathbf{0}$. Nous imposons donc que

$$\int_{D_j} \mathbf{f}_j(x_j) \odot p_j^D(x_j)dx_j = \mathbf{0}, \quad 1 \leq j \leq d.$$

Les contraintes ci-dessus déterminent $\mathbf{f}_0$ comme

$$\mathbf{f}_0 = \int_D \mathbb{E}(\mathbf{Y}|\xi = \mathbf{x}) \odot p^D(\mathbf{x})d\mathbf{x} = (p_0^D)^{-1} \odot \mathbb{E}(\mathbf{Y} \odot \mathbb{I}(\xi \in D)),$$

où $\mathbb{I}(\cdot)$ est la fonction indicatrice. Il est facile de montrer que les fonctions composantes sont alors solutions du système d'équations intégrales

$$\mathbf{f}_j(x_j) = \mathbf{m}_j(x_j) \ominus \mathbf{f}_0 \ominus \bigoplus_{k\neq j} \int_{D_k} \mathbf{f}_k(x_k) \odot \frac{p_{jk}^D(x_j, x_k)}{p_j^D(x_j)} dx_k, \quad 1 \leq j \leq d, \tag{2}$$

où

$$\mathbf{m}_j(x_j) = (p_j^D(x_j))^{-1} \odot \int_{D_{-j}} \mathbb{E}(\mathbf{Y}|\xi = \mathbf{x}) \odot p^D(\mathbf{x})d\mathbf{x}_{-j}$$

et $\ominus$ est la soustraction vectorielle dans $\mathbb{H}$.

Nous construisons alors des solutions approchées des équations intégrales dans (2) sur base de données $(\tilde{\xi}_{i1}, \ldots, \tilde{\xi}_{id}, \tilde{\mathbf{Y}}_i)$, $i = 1, \ldots, n$. Les propriétés non-asymptotiques et asymptotiques des estimateurs obtenus sont alors étudiées.

# 3 Deux exemples

Afin d'illustrer la méthodologie ci-dessous, nous présentons deux exemples. Dans le premier, la variable réponse $\mathbf{Y} \in \mathcal{S}_1^3 = \{(p_1, p_2, p_3)|p_1 + p_2 + p_3 = 1\}$, le simplexe unité de $\mathbb{R}^3$. Dans le second, $\mathbf{Y} \in L^2(S_1^2)$, l'ensemble des fonctions de carrés intégrables sur la sphère unité $S_1^2$ de $\mathbb{R}^3$.

## 3.1 Données compositionnelles

Il est maintenant accepté que les caractéristiques démographiques d'une population et les courants politiques sous-jacents sont des facteurs importants permettant, en partie, de déterminer les résultats d'une élection. Nous illustrons ceci lors des élections présidentielles américaines de 2020 et montrons comment la proportion de personnes possédant un bachelier ($\xi_1$), les revenus par individus ($\xi_2$) et l'âge médian ($\xi_3$) affectent la composition des résultats électoraux dans chaque état. Nous mesurons leur impact sur le vecteur compositionnel $\mathbf{Y} = (Y_1, Y_2, Y_3) \in \mathcal{S}_1^3$, où les $Y_j$ mesurent, respectivement, les proportions de votes pour les démocrates, républicains et autres partis.

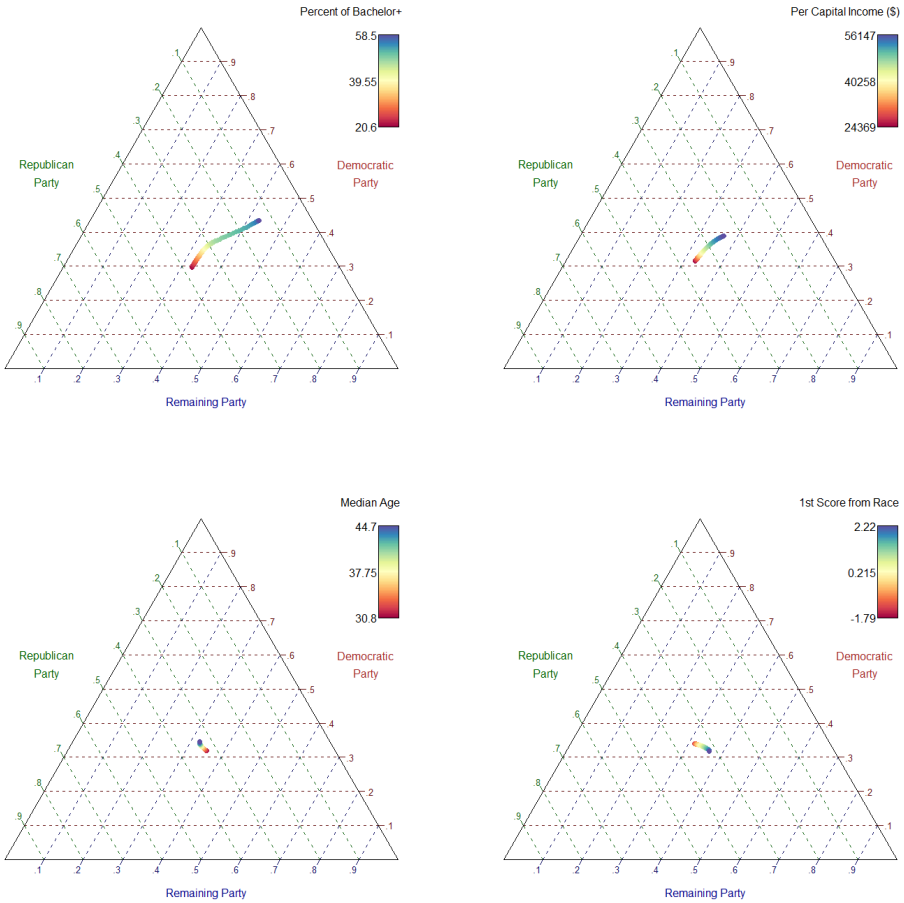



Figure 1: Estimation des fonctions-composantes $\widehat{\mathbf{f}}_j$ pour différentes mesures démographiques par état et leur influence sur les proportions de vote par parti. Pour chaque point dans ces plots ternaires, les proportions pour le parti démocratique, républicain ou autres s'obtiennent en suivant respectivement les lignes rouges, vertes ou bleues.

L'estimation d'un modèle additif pour ce type de données permet d'illustrer l'influence sur les votes de chacun des facteurs $\xi_j$, comme illustré sur la figure ci-dessous.

## 3.2 Données fonctionnelles sur la sphère

Les typhons dévastent chaque année de nombreux pays. Pouvoir prévoir leur itinéraire dès les premières heures de leur existence est crucial. L'administration coréenne météorologique (voir https://data.kma.go.kr/data/typhoonData/typInfoTYList.do?pgmNo=689) maintient une base de données contenant de nombreuses informations sur les typhons apparus en Asie du Sud-Est depuis 2001. Sur base de la pression atmosphérique centrale ($\xi_1$), la vitesse de l'air centrale maximale ($\xi_2$), la vitesse de déplacement ($\xi_3$) et les positions initiales ($\xi_4$), moving direction au temps $T_0$, nous estimons, à l'aide d'un modèle additif, le trajet de divers

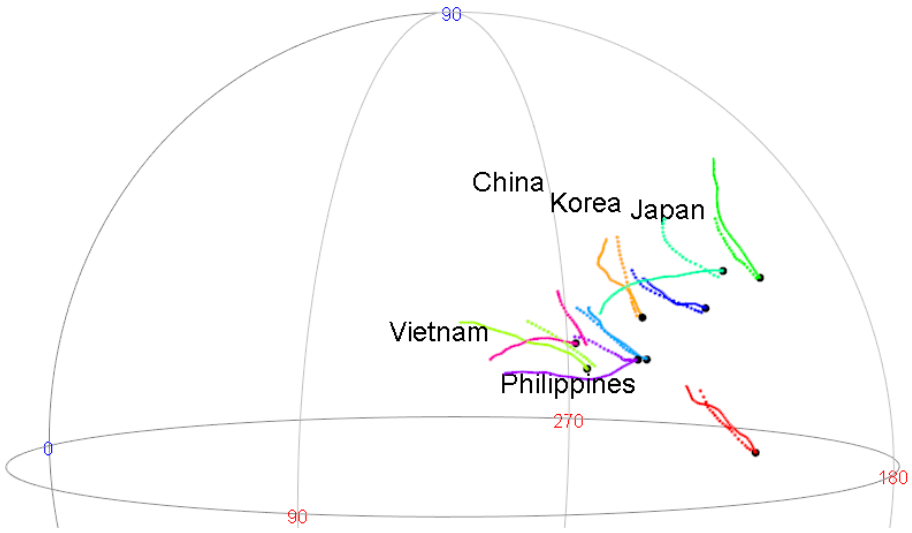



Figure 2: Vraies trajectoires (lignes pleines) et prédites (pointillées) des typhons apparus en 2022.

typhons de l'année 2022 (sur base des $n = 265$ typhons observés de 2001 à 2021).

# Bibliographie

Delaigle, A. (2008). An alternative view of the deconvolution problem, *Statistica Sinica*, 18, pp. 1025-1045.

Mammen, E., Linton, O. B. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics*, 27, pp. 1443-1490.

# Graphes et réseaux

# Régression par processus Gaussiens pour des entrées graphes en grande dimension

Raphaël CARPINTERO PEREZ [1,2] & Sébastien DA VEIGA [3] & Josselin GARNIER [2] & Brian STABER [1]

[1] *Safran Tech, Digital Sciences & Technologies Department, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France,*
*{raphael.carpintero-perez,brian.staber}@safrangroup.com*
[2] *Centre de Mathématiques Appliquées, Ecole Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France,*
*{raphael.carpintero-perez,josselin.garnier}@polytechnique.edu*
[3] *Univ Rennes, Ensai CNRS, CREST - UMR 9194, F-35000 Rennes, France,*
*sebastien.da-veiga@ensai.fr*

**Résumé.** Les algorithmes d'apprentissage statistique appliqués à des données sous formes de graphes ont suscité une grande attention dans des domaines tels que la biochimie, les systèmes de recommandation sociaux et, très récemment, l'apprentissage de simulations basées sur la physique. Les méthodes à noyau, et plus particulièrement la régression par processus Gaussiens, sont particulièrement appréciées car elles sont efficaces lorsque la taille de l'échantillon est faible et lorsqu'il est nécessaire de quantifier les incertitudes de prédiction. Dans cet exposé, nous présentons le noyau entre graphes Sliced Wasserstein Weisfeiler-Lehman (SWWL) qui considère des graphes avec des attributs continus attachés aux sommets. Nous combinons les itérations continues de Wesifeiler Lehman et du transport optimal entre distributions de probabilités empiriques avec la distance de sliced Wasserstein afin de définir une fonction noyau définie positive avec une faible complexité de calcul. Ces deux propriétés permettent de considérer des graphes avec un grand nombre de sommets, ce qui était auparavant une tâche délicate.

**Mots-clés.** Apprentissage statistique, processus Gaussiens, noyaux entre graphes, grande dimension, transport optimal, métamodèles pour la simulation numérique

**Abstract.** Machine learning algorithms applied to graph data have garnered significant attention in fields such as biochemistry, social recommendation systems, and very recently, learning physics-based simulations. Kernel methods, and more specifically Gaussian process regression, are particularly appreciated since they are powerful when the sample size is small, and when uncertainty quantification is needed. In this talk, we introduce the Sliced Wasserstein Weisfeiler-Lehman (SWWL) graph kernel which handles graphs with continuous node attributes. We combine continuous Wesifeiler Lehman iterations and an optimal transport between empirical probability distributions with the sliced Wasserstein distance in order to define a positive definite kernel function with low computational complexity. These two properties make it possible to consider graphs with a large number of nodes, which was previously a tricky task.

**Keywords.** Machine learning, Gaussian processes, graph kernels, high dimension, optimal transport, surrogate models for numerical simulation

# 1 Introduction

Ce travail porte sur la régression par processus Gaussiens indexés par des graphes de grande dimension et possédant des attributs continus.

Dans ce contexte, il est fondamental d'avoir accès à une fonction noyau défini positif entre graphes et dont la complexité est raisonnable afin de pouvoir traiter des graphes en grande dimension (de l'ordre de plusieurs dizaines de milliers de sommets).

Il existe de nombreuses approches qui permettent de définir des noyaux entre graphes, comme l'indiquent les multiples articles de synthèse sur les noyaux entre des graphes (Nikolentzos et al., 2021; Borgwardt et al., 2020). Les approches traditionnelles se concentrent principalement sur la structure d'adjacence des graphes, et peu prennent en compte les possibles *attributs continus* ou sont applicables avec des graphes *de grande taille* et *creux*.

Plus récemment des approches utilisant du transport optimal (Peyré et al., 2019) ont été proposées comme le noyau Wasserstein Weisfeiler Lehman de Togninalli et al. (2019) mais ont une complexité trop élevée et ne permettent pas d'assurer un noyau défini positif. Mais contrairement à la distance de Wasserstein, la distance de sliced Wasserstein (Bonneel et al., 2015) injectée dans des noyaux usuels donne bien des noyaux définis positifs comme l'a mis en avant Meunier et al. (2022) dans le cas de distributions empiriques.

L'approche que nous proposons (Carpintero Perez et al., 2024) est résumée dans la Figure 1. Elle repose sur deux ingrédients: les embeddings de Weisfeiler-Lehman (WL) et la distance de sliced Wasserstein.

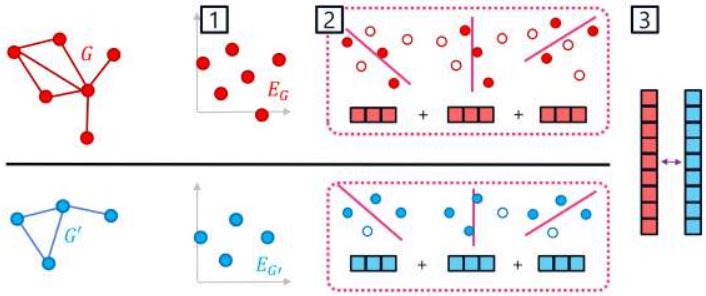



Figure 1: Noyau SWWL. Etape 1: embeddings des graphes. Etape 2: embeddings des quantiles projetés (EQP). Etape 3: distances Euclidiennes entre EQP.

# 2 Notations et présentation du problème

On considère la tâche d'apprentissage d'une fonction $f : \mathcal{G} \to \mathcal{Y}$ où $\mathcal{Y} = \{0, 1\}$ pour les tâches de classification, et $\mathcal{Y} = \mathbb{R}$ pour les tâches de régression. Ici, $\mathcal{G}$ désigne un ensemble de graphes non orientés et éventuellement pondérés ayant des attributs continus. Chaque graphe $G \in \mathcal{G}$ peut donc être représenté comme $G = (V, E, w, \mathbf{F})$ où $V$ est l'ensemble des

sommets et $E$ est un ensemble de paires de sommets, dont les éléments sont appelés arêtes. Les attributs à $d$ dimensions des sommets sont regroupés dans la matrice (de taille $|V| \times d$) $\mathbf{F} = (\mathbf{F}_u)_{u \in V}$. Les poids des arêtes sont attribués par la fonction $w : E \to \mathbb{R}$. Le voisinage d'un sommet $u \in V$ est donné par $\mathcal{N}(u) = \{v \in V : \{u, v\} \in E\}$ et son degré est noté $\deg(u) = |\mathcal{N}(u)|$.

On suppose que l'on dispose d'un ensemble de données $\mathcal{D}$ constitué de $N$ observations $\mathcal{D} = \{(G_i, y_i)\}_{i=1}^N$, où les graphes $G_i$ en entrée peuvent différer en termes de nombres de sommets et de matrices d'adjacence, c'est-à-dire qu'il est possible d'avoir $|V_i| \neq |V_j|$ et/ou $E_i \neq E_j$ pour certains $(i, j) \in \{1, \ldots, N\}^2$. Pour approcher la fonction $f$ on utilise la régression par processus Gaussien. On suppose une loi a priori Gaussienne sur la fonction f (processus Gaussien). Après avoir conditionné selon les observations, la loi a posteriori est à nouveau Gaussienne, ce qui permet d'obtenir les prédictions (grâce à la moyenne) et les incertitudes de prédiction (grâce à la covariance). Ce processus Gaussien est entièrement déterminé par sa moyenne et sa covariance, correspondant à la fonction noyau. On souhaite donc construire une fonction noyau $k : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ définie positive. On renvoie le lecteur à Williams and Rasmussen (2006) pour plus de détails sur la régression par processus Gaussiens.

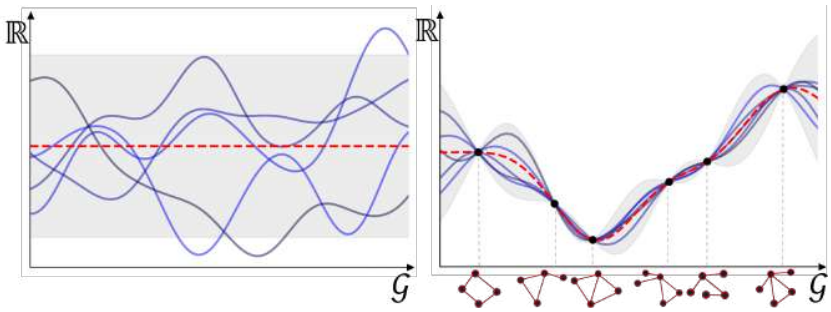



Figure 2: Illustration de la régression par processus Gaussiens pour des entrées de type graphes. Gauche: échantillons de la distribution a priori. Droite: échantillons de la distribution a posteriori après conditionnement sur les observations (les entrées sont des graphes ici).

# 3 Noyau Sliced Wasserstein Weisfeiler-Lehman

La méthodologie proposée repose sur la distance de sliced Wasserstein (Bonneel et al., 2015) qui est basée sur des projections aléatoires et l'expression analytique de la distance de Wasserstein 1D sous forme de quantiles. L'autre ingrédient correspond aux itérations de Weisfeiler-Lehman (WL) continues (Togninalli et al., 2019) qui permettent d'obtenir un embedding des sommets des graphes prenant en compte les attributs continus ainsi que la structure d'adjacence. Les différentes étapes sont détaillées après la définition du noyau.

**Définition 1** (Noyau SWWL)**.** *Soit $P \geqslant 1$ le nombre de projections, $Q \geqslant 2$ le nombre de quantiles, et $H \geqslant 0$ le nombre d'itérations de WL continues. Le noyau SWWL (illustré sur*

*la Figure 1) est défini pour $G, G' \in \mathcal{G}$ par*

$$k_{\mathrm{SWWL}}(G, G') = \exp\left(-\gamma \widehat{SW}^2_{2,P,Q}(\mu_G, \mu_{G'})\right), \tag{1}$$

*où $\mu_G = |V|^{-1} \sum_{u \in V} \delta_{\mathbf{E}^G_u}$ est la mesure empirique associée à l'embedding de WL continu $\mathbf{E}^G = (\mathbf{E}^G_u)_{u \in V}$ de $G$ avec $H$ iterations (voir la Définition 2) et $\gamma > 0$ est un paramètre de précision.*

Le noyau SWWL possède plusieurs propriétés clés pour la régression par processus Gaussiens.

**Propriété 1.** *Il existe une feature map $\phi$ dans un espace de dimension $PQ$ (voir l'Equation (9)) tel que le noyau SWWL peut s'écrire de la façon suivante:*

$$k_{\mathrm{SWWL}}(G, G') = \exp\left(-\gamma ||\phi(\mu_G) - \phi(\mu_{G'})||_2^2\right). \tag{2}$$

L'équation (2) montre que la construction de la matrice de Gram $K = (k(G_i, G_j)_{ij}$ peut se décomposer en deux parties: (1) calcul des embeddings $\phi(\mu_G)$ et (2) assemblage de la matrice de pseudo-distances. On notera que la complexité temporelle de la dernière étape est indépendante des nombres de sommets dans les graphes.

**Propriété 2.** *Soient $\delta$ le degré moyen des sommets et $n$ le nombre moyen de sommets. La complexité temporelle requise pour assembler la matrice de Gram $N \times N$ avec le noyau donné par l'Equation (1) est*

$$\mathcal{O}(NH\delta n + NPn(\log(n) + H) + N^2PQ).$$

**Propriété 3.** *Le noyau SWWL est défini positif.*

On détaille désormais la construction de la feature map $\phi$ de l'Equation (2).

**Embeddings de Weisfeiler-Lehman continus.** Les embeddings de WL ont initialement été proposés pour des graphes ayant des sommets avec des labels discrets, mais ils ont été étendus aux attributs continus par Togninalli et al. (2019). De façon intuitive, les itérations de WL continues mettent à jour les attributs des sommets en agrégeant l'information des voisins. Après $h$ itérations, les embeddings capturent l'information du $h$-voisinage des sommets.

**Définition 2** (Embeddings de WL continus)**.** *Soit $G = (V, E, w, \mathbf{F})$ un graphe possédant les attributs continus $\mathbf{F} = (\mathbf{F}_u)_{u \in V}$, $\mathbf{F}_u \in \mathbb{R}^d$. Les itérations de Weisfeiler-Lehman continues sont définies récursivement pour $h \in \mathbb{N}$ par*

$$\mathbf{F}^{(h+1)}_u = \frac{1}{2}\left(\mathbf{F}^{(h)}_u + \frac{1}{\deg(u)} \sum_{v \in \mathcal{N}(u)} w(u, v)\mathbf{F}^{(h)}_v\right), \tag{3}$$

*avec $\mathbf{F}^{(0)}_u = \mathbf{F}_u$ pour $u \in V$. Etant donné un nombre d'itérations $H \geqslant 0$, l'embedding par sommets de WL continu du graphe $G$ est la concaténation des itérations de WL continues aux étapes $0, 1, \ldots, H$, que l'on note $\mathbf{E}^G = (\mathbf{E}^G_u)_{u \in V}$, $\mathbf{E}^G_u \in \mathbb{R}^{(H+1)d}$.*

**Transport optimal.** Une fois que les embeddings de WL continus $\mathbf{E}^G$ ont été obtenus pour tous les graphes, les distributions empiriques associées peuvent être considérées et les distances de sliced Wasserstein peuvent être calculées entre toutes les paires de distributions empiriques pour construire un noyau.

La distance de sliced Wasserstein (Bonneel et al., 2015) moyenne sur la sphère unité les distances de Wasserstein unidimensionnelles entre les distributions projetées. Cela réduit dans un premier temps la complexité puisque la distance de Wasserstein 1D peut être calculée en $\mathcal{O}(n\log(n))$, et cela donne par ailleurs un noyau de substitution défini positif. La distance de sliced Wasserstein est en effet Hilbertienne (Meunier et al., 2022). On rappelle l'expression de la distance de Wasserstein unidimensionnelle avant de définir la distance de sliced Wasserstein.

**Définition 3** (Distance de Wasserstein unidimensionnelle)**.** *Soit $r \geqslant 1$. La $r$-distance de Wasserstein pour des mesures sur $\mathbb{R}$ a l'expression suivante:*

$$W_r(\mu,\nu) = \left(\int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right|^r dt\right)^{\frac{1}{r}} \tag{4}$$

*où $F_\mu(x) = \mu((-\infty, x]), x \in \mathbb{R}$ est la fonction de répartition et $F_\mu^{-1}(t) = \inf\{x \in \mathbb{R} : F_\mu(x) \geqslant t\}, t \in [0,1]$ est la fonction de répartition inverse (fonction quantile).*

**Définition 4** (Distance de sliced Wasserstein)**.** *Soient $s \geqslant 1$, $r \geqslant 1$. La $r$-distance de sliced Wasserstein est définie par*

$$SW_r(\mu,\nu) := \left(\int_{\mathbb{S}^{s-1}} W_r(\theta_\sharp^*\mu, \theta_\sharp^*\nu)^r d\sigma(\theta)\right)^{\frac{1}{r}}, \tag{5}$$

*où $\mathbb{S}^{s-1}$ est la sphère unité $(s-1)$-dimensionnelle, $\sigma$ est la distribution uniforme sur $\mathbb{S}^{s-1}$, $\theta^* : \mathbf{x} \in \mathbb{R}^s \mapsto \langle \mathbf{x}, \theta\rangle$ la fonction projection dans la direction $\theta \in \mathbb{S}^{s-1}$, $\theta_\sharp^*\mu$ la mesure image de $\mu$ par $\theta^*$, et $W_r$ est la $r$-distance de Wasserstein unidimensionnelle.*

En pratique, une estimation de Monte-Carlo est réalisée en tirant $P$ directions $\theta_1, \ldots, \theta_P$ uniformément dans $\mathbb{S}^{s-1}$.

**Embedding des quantiles projetés.** Le calcul de la distance de Wasserstein 1d entre les mesures empiriques consiste généralement à trier les points projetés, puis à additionner une puissance des distances euclidiennes entre les valeurs aux mêmes rangs. Ici, nous proposons plutôt d'utiliser $Q \ll n$ quantiles équidistants, qui ne dépendent pas des tailles $n$ des distributions empiriques pouvant varier. Cette stratégie diffère des implémentations usuelles (Flamary et al., 2021), où une grille de quantiles est choisie pour chaque paire de distributions en entrée. Plus précisément, soit $0 = t_1 < \cdots < t_\ell < \cdots < t_Q = 1$ une suite de $Q$ points équirépartis sur $[0,1]$, et soit $\theta \in \mathbb{S}^{s-1}$ une direction de projection. L'estimation à $Q$ quantiles de la distance de Wasserstein 1d entre les mesures images $\mu_\theta := \theta_\sharp^*\mu$ et $\nu_\theta := \theta_\sharp^*\nu$ s'écrit

$$\tilde{W}_{r,Q}(\mu_\theta,\nu_\theta) = \left(\frac{1}{Q}\sum_{\ell=1}^{Q}\left|F_{\mu_\theta}^{-1}(t_\ell) - F_{\nu_\theta}^{-1}(t_\ell)\right|^r\right)^{\frac{1}{r}} \tag{6}$$

où $F_\mu(x) = \mu(]-\infty, x])$, $x \in \mathbb{R}$ est la fonction de répartition et $F_\mu^{-1}(t) = \inf\{x \in \mathbb{R} : F_\mu(x) \geqslant t\}$, $t \in [0,1]$ est la fonction de répartition inverse. La distance de sliced Wasserstein donnée par l'Equation (5) est finalement estimée de la façon suivante:

$$\widehat{SW}_{r,P,Q}(\mu,\nu) = \left(\frac{1}{P}\sum_{p=1}^{P} \tilde{W}_{r,Q}(\mu_{\theta_p}, \nu_{\theta_p})^r\right)^{\frac{1}{r}}, \tag{7}$$

où $\theta_1, \ldots, \theta_P$ désignent les $P$ directions de projections tirées uniformément dans $\mathbb{S}^{s-1}$. En combinant les équations (6) et (7), cette estimation peut s'écrire comme

$$\widehat{SW}_{r,P,Q}(\mu,\nu) = ||\phi(\mu) - \phi(\nu)||_r\,, \tag{8}$$

où $||.||_r$ est la $r$-norme dans $\mathbb{R}^{PQ}$, et $\phi$ la feature map explicite

$$\phi_{p+P(q-1)}(\mu) = (PQ)^{-1/r} F^{-1}_{\mu_{\theta_p}}(t_q) \tag{9}$$

pour $p = 1, \ldots, P$ et $q = 1, \ldots, Q$. On appelle cette feature map *embedding des quantiles projetés* (EQP). Il fournit une représentation $PQ$-dimensionnelle de toute distribution de probabilité $\mu$ dans $\mathbb{R}^s$. Dans la définition 1, la distribution de probabilité d'intérêt est la distribution empirique associée aux embeddings de WL continus $\mathbf{E}^G = (\mathbf{E}^G_u)_{u \in V}$ des graphes $G \in \mathcal{G}$ avec $H$ itérations, ce qui donne une dimension $s = H(d+1)$. On obtient de ce fait la propriété 1.

# 4 Expériences

Deux tâches ont été considérées pour tester le noyau SWWL.

Une première concerne la classification de petits graphes avec moins de 100 sommets correspondant à des molécules (Morris et al., 2020) afin de valider les performances du noyau en le comparant à des noyaux de l'état de l'art (en utilisant des séparateurs à vaste marge).

Une seconde concerne la régression par processus Gaussiens pour des graphes provenant de simulations numériques en dynamique et mécanique des fluides basées sur des maillages. La figure 3 montre trois graphes d'entrée provenant du jeu de données Tensile2d [1]. On démontre par ces expériences que le noyau SWWL peut aisément manipuler des graphes à plus de $10^5$ sommets en quelques secondes ou minutes là ou d'autres approches ne passent pas à l'échelle.

Pour ces deux tâches, on obtient des scores de classification/régression équivalents aux méthodes de l'état de l'art comparées mais avec des temps de calculs très inférieurs. On étudie en particulier l'influence des hyperparamètres du modèle: nombre d'itérations, nombre de quantiles et nombre de projections sur les erreurs de prédiction.

[1] `https://plaid-lib.readthedocs.io/en/latest/source/data_challenges/tensile2d.html`

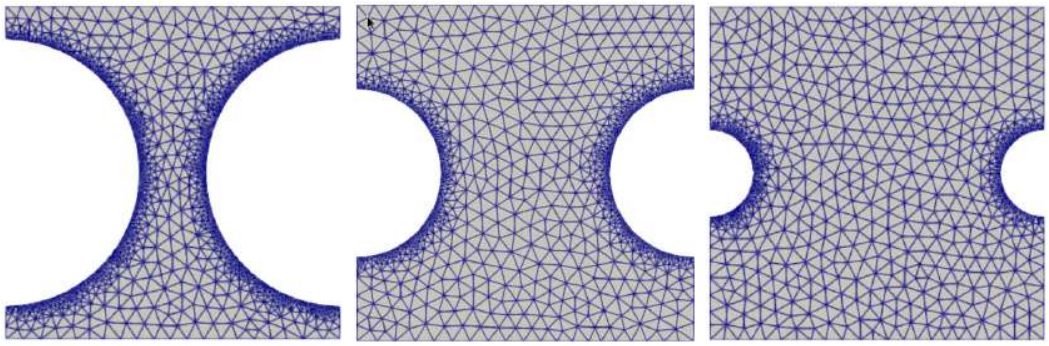


Figure 3: Trois graphes provenant de la version sous-échantillonnée du jeu de données Tensile2d. Les nombres de sommets et les connectivités changent entre les échantillons.

## Remerciements

Ce travail de recherche est mené dans le cadre du projet SAMOURAI (Simulation Analytics and Meta-model-based solutions for Optimization, Uncertainty and Reliability AnalysIs) financé par l'Agence Nationale de la Recherche (ANR-20-CE46-0013).

## References

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., Rieck, B., et al. (2020). Graph kernels: State-of-the-art and future challenges. *Foundations and Trends® in Machine Learning*, 13(5-6):531–712.

Carpintero Perez, R., da Veiga, S., Garnier, J., and Staber, B. (2024). Gaussian process regression with Sliced Wasserstein Weisfeiler-Lehman graph kernels. `https://arxiv.org/abs/2402.03838`.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Meunier, D., Pontil, M., and Ciliberto, C. (2022). Distribution regression with sliced Wasserstein kernels. In *International Conference on Machine Learning*, pages 15501–15523. PMLR.

Morris, C., Kriege, N. M., Bause, F., Kersting, K., Mutzel, P., and Neumann, M. (2020). Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*.

Nikolentzos, G., Siglidis, G., and Vazirgiannis, M. (2021). Graph kernels: A survey. *Journal of Artificial Intelligence Research*, 72:943–1027.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Togninalli, M., Ghisu, E., Llinares-López, F., Rieck, B., and Borgwardt, K. (2019). Wasserstein Weisfeiler-Lehman graph kernels. *Advances in neural information processing systems*, 32.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT press.

# The Deep Latent Position Block Model

Rémi Boutin [1,2] † & Pierre Latouche [2] & Charles Bouveyron [3]2

[1] *Université Paris Cité, CNRS, Laboratoire MAP5, UMR 8145, Paris, France*
[2] *Université Clermont Auvergne, CNRS, Laboratoire LMBP, UMR 6620, Aubière, France*
[3] *Université Côte d'Azur, CNRS, Laboratoire J.A.Dieudonné, INRIA, Maasai team, Nice, France*

**Résumé.**

L'augmentation des capacités de stockage et des données collectées a entraîné un accroissement des jeux de données disponibles. Par conséquent, l'utilisation des réseaux pour modéliser les relations entre différents objets, appelés des nœuds s'est accrue. Ces réseaux pouvant compter un très grand nombre de nœuds, l'information qu'ils contiennent doit-être résumée, le plus souvent à l'aide de méthodes de clustering de nœuds. Afin de rendre les résultats interprétables, une visualisation pertinente du réseau est également requise. Pour ce faire, nous proposons une nouvelle méthodologie appelée Deep-LPBM, permettant d'obtenir simultanément un clustering des noeuds basé sur une approche par bloc, plus générale que la détection de communautés, ainsi qu'une représentation continue des nœuds dans un espace latent. Deep-LPBM utilise une stratégie d'auto encodeur variationnel, s'appuyant sur un réseau de convolution de graphe, avec un décodeur adapté. L'inférence repose sur la vraisemblance marginale du modèle, et l'optimisation alterne entre des équations analytiques ainsi qu'une descente de gradient stochastique. Ce travail étant en cours, des expériences sur données simulées ainsi que sur données réelles seront fournies si le papier est accepté pour une présentation orale.

**Mots-clés.** Clustering de nœuds, auto-encodeur de graphe variationnel, modélisation par blocs, visualisation de graphe

**Abstract.** The increase in the quantity of data has led to a soaring use of networks to model relationships between different objects, called nodes. Since the number of nodes can be very large, the network information must be summarised, mostly with node clustering methods. In order to make the results interpretable, a relevant visualization of the network is also required. To tackle those two issues, we propose a new method called Deep-LPBM which provides simultaneously a network visualization based on block modelling, allowing a more general clustering than community detections, as well as a continuous representation of nodes in a latent space. Our methodology is based on a variational autoencoder strategy, relying on a graph convolutional network, with a specifically designed decoder. The inference is based on the marginal likelihood of the model, and the optimisation combines analytical equations with stochastic gradient descent. As this work is ongoing, experiments on simulated data as well as on real data will be provided if the paper is accepted for an oral presentation.

**Keywords.** Node clustering, variational graph auto-encoder, block modelling, network visualisation

†mail: remi.boutin.stat@gmail.com

# 1 Introduction and contribution

Networks are encountered in a variety of fields, ranging from social sciences to biology. Their capacity to represent any type of object and relationship makes it a core object to model interactions. However, they present difficulties to apprehend since non-observed features, such as node cluster memberships, may impact the observed network topology and engender specific connectivity patterns. This requires the development of specifically fashioned methods, models and inference strategies to capture information. To give an example, one of the most studied type of groups is called a community and corresponds to nodes highly connected to nodes of the same group but poorly connected to nodes from other groups. The community-detection methods are numerous but do not necessarily generalise to other types of structure such as a star pattern (a cluster of nodes poorly connected together but highly connected to nodes from the other clusters). Therefore, being able to capture the structure underlying the data is essential to model any type of relationship, even without prior knowledge of the network topology. This flexibility was provided by the stochastic block model (SBM, Snijders and Nowicki, 1997; Daudin, Picard, et al., 2008), enabling to model any type of connectivity patterns among the network. However, this flexibility comes at the cost of representation. Indeed, those methodologies do not provide a direct representation of the network, but only a high-level depiction of the underlying patterns, where clusters are represented as nodes, cluster sizes as node sizes, and the number of connections between clusters as edge widths. This meta representation may hide some node-specific properties. For instance, a node in a cluster may be more connected to another group than the other nodes in its cluster. Hence, this feature should appear in the network representation, since this node might play a crucial role in the network, precisely because it connects two clusters. To represent a network, positional approaches have been proposed by performing link prediction based on the similarity between estimated continuous node representations (Hoff et al., 2002; Handcock et al., 2007). Unfortunately, such methods only estimate communities. While some previous work has focused on combining the two (Hoff, 2007; Daudin, Pierre, et al., 2010), the methodology we propose is the first to incorporate the graph neural network ability to create informative node embeddings. To this end, a marginalised block model is introduced, where a logistic-Gaussian distribution models the node cluster membership probabilities and is used for visualisation purposes. This is an ongoing work that will come with a Python package as well as an extensive benchmark and a real world use-case.

**Notations** In this paper, matrices and collections of vectors are denoted in bold cases $\mathbf{X}$, the space of $n \times m$ matrices with coefficient in $E$ is denoted $\mathcal{M}_{n\times m}(E)$, and should not be confused with the multinomial distribution denoted $\mathcal{M}_n(m, p)$ where $n$ is the dimension of the vector, $m$ is the number of draws and $p = (p_1, \dots, p_n) \in \Delta_n$ is the probability vector. The $n$-dimensional simplex is denoted $\Delta_n = \{p \in \mathbb{R}^n : \forall i, p_i \geq 0 \text{ and } \sum_{i=1}^n p_i = 1\}$. Moreover, the network is denoted $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \dots, N\}$ denotes the set of vertices and $\mathcal{E}$ the set of edges, and the binary adjacency matrix $\mathbf{A} \in \mathcal{M}_{N\times N}(\{0, 1\})$ such that $A_{ij} = 1$ if and only if $(i, j) \in \mathcal{E}$. In this work, the graph is assumed to be directed, meaning that $\mathbf{A}$ does not need to be symmetric.

## 2 Assumptions regarding the network generation

In this section, we present the assumptions concerning the graph generation. Assuming that the number of clusters $Q$ is fixed beforehand, each node $i \in \mathcal{V}$ is assumed to belong to a cluster, represented by the cluster membership vector $C_i$. The variables $(C_i)_i$ are assumed to be independent and identically distributed (i.i.d) according to a multinomial distribution such that:

$$C_i \overset{i.i.d}{\sim} \mathcal{M}_Q(1, \gamma),$$

with $\gamma \in \Delta_Q$ the vector of cluster proportions. The vectors $C_i \in \{0,1\}^Q$ are one-hot encoded, with $C_{iq} = 1$ if node $i$ belongs to cluster $q$ and $C_{iq} = 0$ otherwise. The probability of the cluster membership matrix $\mathbf{C} = (C_1, \dots, C_N)^T \in \mathcal{M}_{N\times Q}(\{0,1\})$ is given by:

$$p(\mathbf{C} \mid \gamma) = \prod_{i=1}^{N} \prod_{q=1}^{Q} \gamma_q^{C_{iq}}. \tag{1}$$

Given the cluster membership matrix $\mathbf{C}$, the nodes are assumed to be independent, and represented by a Gaussian vector $Z_i$ in a $Q-1$ dimensional latent space:

$$Z_i \mid C_{iq} = 1 \overset{i.i.d}{\sim} \mathcal{N}_{Q-1}\left(\mu_q, \sigma_q^2 \mathbf{I}_{Q-1}\right). \tag{2}$$

The set of node embeddings is denoted $\mathbf{Z} = (Z_i)_i$ in the rest of the paper, and the set of means and variances are denoted $\boldsymbol{\mu} = (\mu_q)_q$ and $\boldsymbol{\sigma}^2 = (\sigma_q^2)_q$ respectively. To link the latent representations of the nodes $Z_i$, with the block modelling, we rely on the *bijective softmax* transformation, as presented in Xu et al. (2014), $h\colon Z_i \in \mathbb{R}^{Q-1} \mapsto \eta_i \in \Delta_Q$ where:

$$\eta_{iq} = \begin{cases} e^{Z_{iq}} / \left(1 + \sum_{r=1}^{Q-1} e^{Z_{ir}}\right) & \text{if } q \in \{1, \dots, Q-1\} \\ 1 / \left(1 + \sum_{r=1}^{Q-1} e^{Z_{ir}}\right) & \text{if } q = Q \end{cases}, \tag{3}$$

and we denote $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)^\top \in \mathcal{M}_{N\times Q}\big((0,1)\big)$. The mapping $h$ aims at encoding the $(Z_i)_i$ into cluster membership probabilities. Eventually, the probability of connection between two nodes follows a Bernoulli distribution with parameters depending on $\eta$ such that:

$$p(\mathbf{A} \mid \mathbf{Z}, \mathbf{\Pi}) = \prod_{i\neq j} p(A_{ij} \mid Z_i, Z_j, \mathbf{\Pi}) = \prod_{i\neq j} \left(\eta_i^\top \mathbf{\Pi} \eta_j\right)^{A_{ij}} \left(1 - \eta_i^\top \mathbf{\Pi} \eta_j\right)^{1-A_{ij}}, \tag{4}$$

where $\mathbf{\Pi} = (\pi_{qr})_{1\leq q,r\leq Q} \in \mathcal{M}_{Q\times Q}\big((0,1)\big)$ is the matrix of probability of connection between clusters. Consequently, the joint distribution of $(\mathbf{A}, \mathbf{Z}, \mathbf{C})$ can be factorised as:

$$p(\mathbf{A}, \mathbf{Z}, \mathbf{C} \mid \mathbf{\Pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}) = p(\mathbf{A} \mid \mathbf{Z}, \mathbf{\Pi}) p(\mathbf{Z} \mid \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\mathbf{C} \mid \boldsymbol{\gamma}). \tag{5}$$

## 3 Inference

The next section presents the inference as well as the optimisation.

## 3.1 Likelihood

To estimate the parameters, we rely on the marginal likelihood of the network, with latent variables $\mathbf{C}$ and $\mathbf{Z}$, and the set of parameters $\Theta = \{\mathbf{\Pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}\}$. From Equations (1), (2) and (4), we can deduce that the marginal log-likelihood is given by:

$$\mathcal{L}(\Theta; \mathbf{A}) = \log p(\mathbf{A} \mid \Theta) = \log \left( \sum_{\mathbf{C}} \int_{\mathbf{Z}} p(\mathbf{A}, \mathbf{C}, \mathbf{Z} \mid \Theta) d\mathbf{Z} \right). \tag{6}$$

Unfortunately, this quantity is not tractable since the sum over $\mathbf{C}$ requires to compute $Q^N$ terms. Therefore, we choose to rely on a variational inference strategy for approximation purposes.

**Decomposition of the marginal log-likelihood** For any distribution $R(\mathbf{C}, \mathbf{Z})$, the following decomposition holds:

$$\mathcal{L}(\Theta; \mathbf{A}) = \mathscr{L}(R(\cdot); \Theta) + \mathrm{KL}\left(R(\cdot) || p(\mathbf{C}, \mathbf{Z} \mid \mathbf{A})\right), \tag{7}$$

where the expected lower bound (ELBO) is given by:

$$\mathscr{L}(R(\cdot); \Theta) = \mathbb{E}_R \left[ \log \frac{p(\mathbf{A}, \mathbf{C}, \mathbf{Z} \mid \Theta)}{R(\mathbf{C}, \mathbf{Z})} \right]. \tag{8}$$

Since the Kullback-Leibler divergence is always positive in Equation (7), the ELBO is a lower bound of the marginal log-likelihood. Since the marginal log-likelihood does not depend on $R(\cdot)$, maximizing the ELBO with respect to $R(\cdot)$ is equivalent to minimizing the Kullback-Leibler divergence between $R(\cdot)$ and the posterior distribution. Hence, we restrict the family of variational distributions by considering a mean-field assumption as well as the following hypotheses to make the ELBO tractable:

$$R(\mathbf{C}, \mathbf{Z} \mid \mathbf{A}) = R(\mathbf{C}) R(\mathbf{Z} \mid \mathbf{A}), \tag{9}$$

$$R(\mathbf{C}) = \prod_{i=1}^{N} R_{\tau_i}(C_i) = \prod_{i=1}^{N} \mathcal{M}_Q(C_i; 1, \tau_i), \tag{10}$$

$$R(\mathbf{Z} \mid \mathbf{A}) = \prod_{i=1}^{N} R_\phi(Z_i \mid \mathbf{A}) = \prod_{i=1}^{N} \mathcal{N}_{Q-1}(Z_i; \mu_\phi(\mathbf{A})_i, \sigma^2_\phi(\mathbf{A})_i \mathbf{I}_{Q-1}), \tag{11}$$

where $\boldsymbol{\tau} = (\tau_i)_{i=1}^N$ with $\forall i \in \{1, \dots, N\}$, $\tau_i \in \Delta_Q$. Moreover, in Equation (11), the mapping $\mu_\phi : \mathcal{M}_{N \times N}(\mathbb{R}) \mapsto \mathcal{M}_{N \times (Q-1)}(\mathbb{R})$ ($\sigma^2_\phi : \mathcal{M}_{N \times N}(\mathbb{R}) \mapsto (\mathbb{R}^+)^N$ respectively) is the mapping normalising the adjacency matrix (with 1 on its diagonal for numeric stability) by its degree, $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}(\mathbf{A} + \mathbf{I}_N)\mathbf{D}^{-1/2}$, and encoding the normalised adjacency matrix into the approximated posterior means (standard deviations) of the node latent positions. The diagonal matrix $\mathbf{D}$ is filled with $D_{ii} = \sum_{j=1}^{N} (\mathbf{A} + \mathbf{I}_N)_{ij}$. The two mappings $\mu_\phi$ and $\sigma^2_\phi$ rely on a GCN parametrised by $\phi$. Regarding the encoder of the adjacency matrix, we based our neural network architecture on Kipf and Welling (2016). Thus, the ELBO can be decomposed as follows:

$$
\begin{aligned}
\mathscr{L}(R(\cdot);\Theta) =& \mathbb{E}_R\left[\log p(\mathbf{A} \mid \mathbf{Z}, \mathbf{\Pi})\right] + \mathbb{E}_R\left[\log p(\mathbf{Z} \mid \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\sigma})\right] - \mathbb{E}_R\left[\log R(\mathbf{Z} \mid \mathbf{A})\right] \\
&+ \mathbb{E}_R\left[\log p(\mathbf{C} \mid \boldsymbol{\gamma})\right] - \mathbb{E}_R\left[\log R(\mathbf{C})\right]. \\
=& \underbrace{\mathbb{E}_R\left[\log p(\mathbf{A} \mid \mathbf{Z}, \mathbf{\Pi})\right]}_{\textbf{Reconstruction loss}} \underbrace{- \mathrm{KL}(R(\mathbf{Z} \mid \mathbf{A}) || p(\mathbf{Z} \mid \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\sigma})) - \mathrm{KL}(R(\mathbf{C}) || p(\mathbf{C} \mid \boldsymbol{\gamma}))}_{\textbf{Regularising term}} \\
=& \sum_{i,j=1}^{N} \Big( A_{ij} \mathbb{E}_R\left[\log \eta_i^\top \mathbf{\Pi} \eta_j\right] + (1 - A_{ij}) \mathbb{E}_R\left[\log\left(1 - \eta_i^\top \mathbf{\Pi} \eta_j\right)\right] \Big) \\
&- \sum_{i=1}^{N} \sum_{q=1}^{Q} \tau_{iq} \, \mathrm{KL}_{iq}\left(\mu_\phi(\mathbf{A})_i, \sigma_\phi(\mathbf{A})_i, \mu_q, \sigma_q\right) - \sum_{i=1}^{N} \sum_{q=1}^{Q} \tau_{iq} \log \frac{\tau_{iq}}{\gamma_q},
\end{aligned}
\tag{12}
$$

where

$$
\mathrm{KL}_{iq}\left(\mu_\phi(\mathbf{A})_i, \sigma_\phi(\mathbf{A})_i, \mu_q, \sigma_q\right) = \log \frac{\sigma_q^{(Q-1)}}{\sigma_\phi(\mathbf{A})_i^{(Q-1)}} - \frac{Q-1}{2} + \frac{(Q-1)\sigma_\phi^2(\mathbf{A})_i + \|\mu_\phi(\mathbf{A})_i - \mu_q\|_2^2}{2\sigma_q^2}.
$$

## 3.2 Optimisation

To optimise the ELBO, we propose to alternate between closed-form updates and stochastic gradient descent steps thanks to the results presented in the next section.

**Analytical updates with respect to the model parameters $\boldsymbol{\tau}$, $\gamma$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$** The first-order conditions applied to the ELBO with respect to $\boldsymbol{\tau}$ and the model parameters give closed-form updates as stated in the following proposition.

**Proposition 1.** *Let $\mathscr{L}(R(\cdot);\Theta)$ denote the ELBO described in Equation* (12)*. The first-order conditions with respect to $\tau$, $\gamma$, $(\mu_q)_q$ and $(\sigma_q)_q$ give the following updates:*

$$
\tau_{iq} = \frac{\gamma_q e^{-\mathrm{KL}_{iq}}}{\sum_{r=1}^{Q} \gamma_r e^{-\mathrm{KL}_{ir}}}, \tag{13}
$$

$$
\gamma_q = \frac{1}{N} \sum_{i=1}^{N} \tau_{iq}, \tag{14}
$$

$$
\mu_q = \left( \sum_{i=1}^{N} \tau_{iq} \right)^{-1} \sum_{i=1}^{N} \tau_{iq} \mu_\phi(\mathbf{A})_i, \tag{15}
$$

$$
\sigma_q^2 = \left( (Q-1) \sum_{i=1}^{N} \tau_{iq} \right)^{-1} \sum_{i=1}^{N} \tau_{iq} \left( (Q-1)\sigma_\phi^2(\mathbf{A})_i + \|\mu_\phi(\mathbf{A})_i - \mu_q\|_2^2 \right). \tag{16}
$$

One way to interpret the update of $\boldsymbol{\tau}$ is to note that the optimal probability of node cluster membership with respect to cluster $q$ decreases exponentially fast with the Kullback-Leibler divergence between the variational distribution of $Z_i$ and the distribution of the

representation of cluster $q$. The update of $\gamma$ corresponds to the approximated posterior expectation of the cluster proportions in the network. On the one hand, the optimal $\mu_q$ is given by the posterior means of $(Z_i)_i$ weighted by each node contribution to the corresponding cluster. On the other hand, the optimal variance $\sigma_q^2$ is given by the sum of two terms. The first one corresponds to the weighted mean of the posterior variances of the nodes. The second one corresponds to the weighted mean of the squared Euclidean distances between the node posterior means and $\mu_q$. In other words, the variances incorporate both the uncertainty about the posterior variance, illustrated by the $\sigma_\phi^2(A)_i$ terms, as well as the uncertainty regarding the $\mu_\phi$, corresponding to the $\|\mu_\phi(\mathbf{A})_i - \mu_q\|_2^2$ terms.

**Stochastic gradient descent** One of the core difficulties in this model is the estimation of the parameters $\mathbf{\Pi}$ and the variational parameters $\phi$ due to the intractable term $\mathbb{E}_R\left[\log p(A \mid \mathbf{Z}, \mathbf{\Pi})\right]$, and in particular $\mathbb{E}_R\left[\log(\sum_{q,r} \eta_{i,q}\eta_{j,r}\pi_{qr})\right]$. To overcome this issue, we rely on a stochastic gradient descent algorithm using the reparametrisation trick (Kingma and Welling, 2014; Rezende et al., 2014) enabling easy computations of the gradient estimates with low variances.

# 4 Evaluation on synthetic datasets

In real-life datasets, quantifying the relevance of a network representation as well as node partitions is a challenging task since no partition of the node exists. Therefore, to assess Deep-LPBM ability to cluster the data, it is necessary to compare its clustering results with a ground truth on synthetic data. First, we present the network structures used in this section to evaluate our methodology. Second, we illustrate the information provided by Deep-LPBM results on the challenging disassortative structures. Third, a comparison of Deep-LPBM representational capacity with the variational graph auto-encoder (VGAE, Kipf and Welling, 2016) and the deep latent position model (Liang et al., 2022:Deep-LPM, ) on the three proposed connectivity structures is exposed. To end this section, we give a quantitative assessment of the clustering results on the community, the hub and the disassortative structures and compare our results with Deep-LPM clustering and the K-means algorithm applied on VGAE embeddings.

## 4.1 Presentation of network structures and Deep-LPBM results

To assess Deep-LPBM capacity to represent different network topologies, we evaluate our methodology on three different structures, all composed of 100 nodes and 5 clusters. The **community structure**, where nodes in the same cluster have a high probability of connection set to 0.5, while nodes in different clusters have a low probability of connection set to 0.01. The **disassortative structure** where nodes in different cluster have a high probability of connection set to 0.5, while nodes in the same cluster have a low probability of connection set to 0.01. Finally, we also use a **hub structure** which is a combination of both with one cluster following the disassortative pattern and the four others being communities.

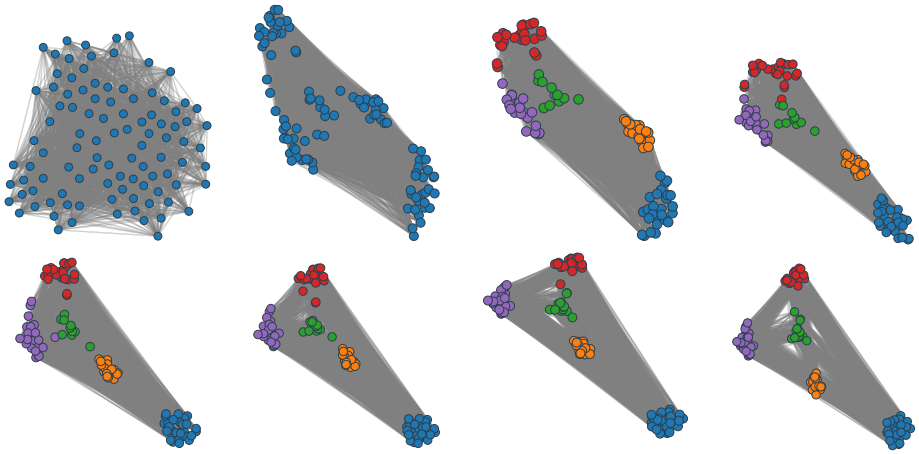


Figure 1: Evolution of the node embeddings during the estimation of Deep-LPBM on a disassortative network. The networks at the top (at the bottom respectively) correspond, from left to right, to the embeddings at the start of the GCN initialisation, the end of the GCN initialisation, iteration 100 and iteration 200 of Deep-LPBM (iteration 500, 1000, 1500, 2000 respectively). The embeddings were projected in $\mathbb{R}^2$ using the t-sne algorithm.

#### 4.1.1 Learning node representations

These two sections highlight the flexibility of Deep-LPBM by analysing both the block modelling estimates as well as the network representation. We start with the latter, with the evolution of the representation during the optimisation presented in Figure 1.

The results provided in this section are obtained by fitting Deep-LPBM on a disassortative network and projecting the estimated embeddings in $\mathbb{R}^2$ with a t-sne algorithm (Van der Maaten and Hinton, 2008). First, we observe an efficient separation of the clusters which cannot be obtained with a similarity-based decoder since the probability of connection would increase with the correlation of the node embeddings. Therefore, the latent space would not be able to show any structure, as depicted in Figure 3. On the contrary, the model we propose is capable of imposing a structure on the variational distribution such as to obtain a node latent space matching with the connectivity patterns of the network.

#### 4.1.2 Block modelling information

Although positional models offer a visualisation of the entire network, meta-representation of a network, such as Figure 2 can only be obtained by a block modelling strategy. The connectivity structure of the graph, captured by the matrix $\mathbf{\Pi}$, is displayed in Figure 2 as well as the associated meta-graph. A meta-graph is a network composed of nodes representing the estimated clusters with a size proportional to the estimated cluster proportions $\gamma$. The edge widths of the meta-graph are proportional to $\mathbf{\Pi}$. In large networks, this type of representation presents the advantage of being easily interpretable as well as focusing on the generative understanding of the network. Here, it is clear that the clusters are highly connected one to another but nodes from the same cluster are poorly connected together.

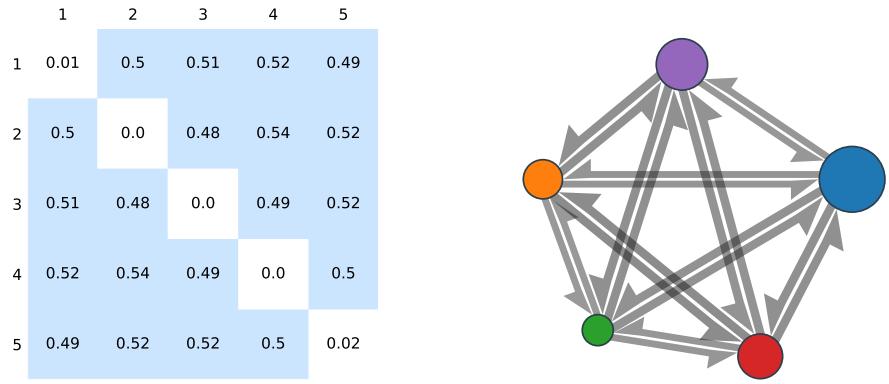



Figure 2: Meta-network based on Deep-LPBM results with an underlying disassortative structure. On the left-hand side, we provide the estimation of the connection probability matrix $\mathbf{\Pi}$. On the right-hand side, the meta-network is composed of nodes representing the clusters, their size is proportional to the corresponding estimated cluster proportion $\gamma$ and the edge widths are proportional to $\mathbf{\Pi}$. A threshold has been set such that edges corresponding to a probability of connection lesser or equal to 0.02 are not displayed.

## 4.2 Representational power on three network structures

In Figure 3, both VGAE and Deep-LPBM efficiently render a latent space capturing the community structure as well as the hub structure. Eventually, the disassortative structure is more difficult to represent. Let us recall that the VGAE decoder models the probability of connection between two nodes with a sigmoid function applied to the cosine similarity between their respective latent positions. Therefore, it necessarily fails to capture the disassortative structure of the network. Hence, two nodes in the same cluster, that have a low probability of connection cannot be represented similarly, and thus cannot be close in the latent space. Conversely, Deep-LPBM is able to translate the connectivity pattern into the position of the nodes, such that nodes of the same cluster, poorly connected together, are close in the latent space.

## 4.3 Clustering evaluation on synthetic data

In this section, we aim to assess the clustering performance of our methodology. We compare it with Deep-LPM that relies on node embedding similarity as a decoder as well as VGAE used to estimate the node embeddings followed by K-means algorithm fitted on these embeddings. The benchmark is performed on networks with 100 nodes and $Q = 5$ clusters. The results are displayed in Table 1. First, we note the efficiency of Deep-LPBM on communities and hubs, reaching an ARI of 1 in both cases. It does not degrade the good performance of its competitors on these architectures, with ARIs of 1 and 0.97 for the VGAE and K-means, and an ARI of 1 on both structures for Deep-LPM. However, these competitors are not able to represent any connectivity structure in the disassortative case. In particular, as shown in Figure 3 for the VGAE, they cannot find relevant node clusters in this setting and end up with an ARI of 0.02 and 0. On the contrary, our methodology reaches an ARI of 0.8, much

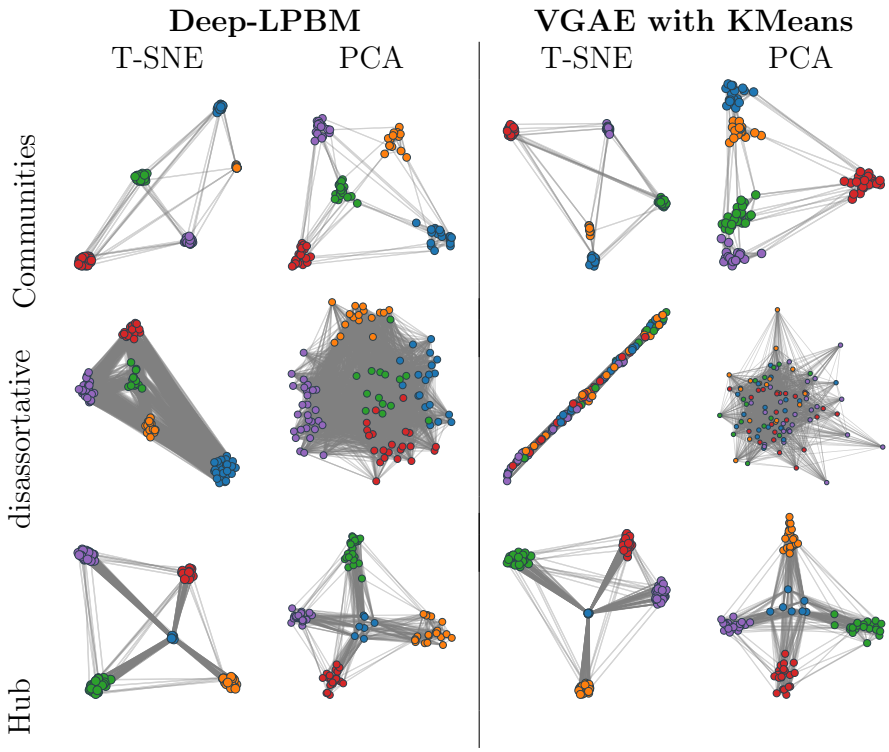



Figure 3: Example of three networks with a latent structure composed of (top to bottom) five communities, a disassortative network with five clusters and four communities with a hub. On the left-hand side (right-hand side respectively), the networks represent the results obtained fitting Deep-LPBM (VGAE) and using a t-sne projection (left) as well as a PCA (right). Each node colour corresponds to its true cluster membership.

higher than its competitors.

# 5 Conclusions

This paper introduced a new methodology combining a block model with a deep latent position model. By modifying the edge distribution and marginalising over a bijective transformation of the node latent representations, we managed to use the node embeddings as cluster probability memberships. We obtained richer results, providing a high-level meta-network, as well as a full network representation, to incorporate details at the node-level. Deep-LPBM is based on the encoder of a graph variational autoencoder combined with a novel block model decoder. Experiments showed that on communities, hubs and disassortative networks, our methodology rightfully translated the network salient information into the latent space. In addition, the clustering results are competitive with the state-of-the-art Deep-LPM. This is an on going work and a more extensive benchmark will be provided if the paper were to be accepted.

| | Communities | disassortative | Hub |
|---|---|---|---|
| VGAE + Kmeans | $1.00 \pm 0.01$ | $0.02 \pm 0.02$ | $0.97 \pm 0.06$ |
| Deep-LPM | $1.00 \pm 0.00$ | $0.00 \pm 0.00$ | $1.00 \pm 0.00$ |
| **Deep LPBM** | $1.00 \pm 0.00$ | $0.80 \pm 0.08$ | $1.00 \pm 0.00$ |

Table 1: ARI obtained by a K-means algorithm applied on VGAE node embeddings, Deep-LPM and Deep-LPBM partitions. We keep the best initialisation ELBO wise over 10 initialisations and repeat it over 10 network to obtain the means and standard deviations.

# References


Daudin, J-J, Franck Picard, and Stéphane Robin (2008). *A mixture model for random graphs*. In: Statistics and computing, Vol. 18, No. 2, pp. 173–183.

Daudin, Jean-Jacques, Laurent Pierre, and Corinne Vacher (2010). *Model for heterogeneous random networks using continuous latent variables and an application to a tree–fungus network*. In: Biometrics, Vol. 66, No. 4, pp. 1043–1051.

Handcock, Mark S, Adrian E Raftery, and Jeremy M Tantrum (2007). *Model-based clustering for social networks*. In: Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 170, No. 2, pp. 301–354.

Hoff, Peter (2007). *Modeling homophily and stochastic equivalence in symmetric relational data*. In: Advances in neural information processing systems, Vol. 20, pp. 657–664.

Hoff, Peter D, Adrian E Raftery, and Mark S Handcock (2002). *Latent space approaches to social network analysis*. In: Journal of the american Statistical association, Vol. 97, No. 460, pp. 1090–1098.

Kingma, Diederik P and Max Welling (2014). *Auto-Encoding Variational Bayes*. arXiv: 1312.6114 [stat.ML].

Kipf, Thomas N and Max Welling (2016). *Variational graph auto-encoders*. arXiv: 1611.07308 [stat.ML].

Liang, Dingge et al. (2022). *Deep latent position model for node clustering in graphs*. In: The 30th European Symposium on Artificial Neural Networks (ESANN 2022).

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). *Stochastic backpropagation and approximate inference in deep generative models*. In: International conference on machine learning. Proceedings of Machine Learning Research, pp. 1278–1286.

Snijders, Tom AB and Krzysztof Nowicki (1997). *Estimation and prediction for stochastic blockmodels for graphs with latent block structure*. In: Journal of classification, Vol. 14, No. 1, pp. 75–100.

Van der Maaten, Laurens and Geoffrey Hinton (2008). *Visualizing data using t-SNE*. In: Journal of machine learning research, Vol. 9, No. 86, pp. 2579–2605.

Xu, Zhiqiang, Yiping Ke, and Yi Wang (2014). *A Fast Inference Algorithm for Stochastic Blockmodel*. In: 2014 IEEE International Conference on Data Mining, pp. 620–629.

# Histogram-based approach for graphon estimation via joint exploitation of multiple networks

**Roland B. Sogan and Tabea Rebafka**

*Sorbonne Université, Université Paris Cité, CNRS*
*Laboratoire de Probabilités, Statistique et Modélisation*
*4 place Jussieu, 75005 Paris, France*
*Email: roland-boniface.sogan@sorbonne-universite.fr,*
*tabea.rebafka@sorbonne-universite.fr*

February 13, 2024

**Abstract**

Exchangeable graph models represent a commonly used non-parametric approach in the modeling of network data. They are characterized by a mathematical object called a graphon. This study focuses on estimating the graphon from multiple networks generated independently from the same model, with possibly distinct sets of nodes. We propose a new histogram estimator that leverages the joint sorting of empirical degrees of the graphs. Our algorithm is extremely fast and scalable to very huge datasets. A numerical study illustrates that the proposed estimator clearly outperforms naive estimators based on the average of individual graphon estimates. Our estimator is consistent when the number of nodes per network increases.

*Keywords:* Graphon, random graphs, multiple networks, histogram estimate.

## 1 Introduction

Machine learning on network-valued data is an active research field. For a long time research has focused on the analysis of a single network, but today in more and more applications entire sets of networks are observed. Thus, there is an increasing interest in the joint study of multiple networks. In this work, we consider the important problem of recovering the graphon from which a collection of graphs was generated. A graphon is a bi-variate function that represents the limiting structure of a sequence of graphs with increasing number of nodes (Lovász and Szegedy, 2006), but it can also be viewed as a nonparametric statistical model for exchangeable random graphs (Diaconis and Janson, 2007; Bickel and Chen, 2009). In general, the graphon is not uniquely defined, but only up to permutation of its parts. In some cases, a canonical representation of the graphon can be easily defined.

The focus of this work is on sets of networks without node correspondence, that is, every network comes with its own set of vertices. The lack of node correspondence represents a challenge for the analysis of the data, as in every graph node labels are arbitrary and can be

permuted in any way. Thus, comparing networks becomes tricky and cannot be done in a straightforward way. For instance, the Euclidean distance between two adjacency matrices is meaningless as it depends on the order of the nodes. Moreover, as node sets are different from one network to another, networks may also have different numbers of nodes, so that adjacency matrices are of different sizes and not comparable at all. Thus, the joint analysis of multiple networks without node correspondence is a challenge in general and in particular when it comes to the estimation of the graphon. To the best of our knowledge, the only graphon estimator for multiple networks that have different sets of nodes provided by the literature is based on the stochastic block model (SBM), which is a model with a specific piecewise constant graphon. That is, the observed networks are considered as independent and identically distributed (iid) realizations of a SBM. For fitting a SBM to a collection of networks, Chabert-Liddell et al. (2023) propose a variational EM-algorithm, while Rebafka (2024) proposes a hierarchical greedy algorithm. Both methods have long computing time and are not scalable to very huge datasets. The purpose of this work is to propose and discuss new methods for the fast estimation of the graphon in the multiple network setting.

In the literature of machine learning, several graphon estimation procedures have been proposed based on the observation of a single network or multiple networks with node correspondence, that is, the observed graphs share the same node set. Besides the above mentioned SBM, Airoldi et al. (2013) propose an estimation procedure for a graphon, called the stochastic blockmodel approximation (SBA) algorithm, computed on multiple networks with node correspondence. The basic idea of SBA is to approximate the graphon by a two-dimensional step function, which corresponds to a SBM. Now, SBA is a fast greedy algorithm, where the number of blocks is chosen in a data-driven way. So, when the number of nodes increases, a SBM with more blocks is selected yielding a step-function approximation of the graphon with ever finer intervals. As a consequence consistency of the SBA algorithm is obtained. Another approach is the universal singular value thresholding (USVT) algorithm for the observation of a single network (Chatterjee, 2012), which is based on a SVD of the adjacency matrix. The estimator is shown to be consistent when the number of nodes tends to infinity. Moreover, Lloyd et al. (2012) introduce a Bayesian nonparametric model by placing a Gaussian process prior on the graphon. However, there is no consistency guarantee of the estimator. None of these graphon estimators has canonical form. However, Chan and Airoldi (2014) propose a consistent and numerically efficient histogram estimator of the canonical form of the graphon. Their estimator is based on the so-called sorting and smoothing (SAS) algorithm, which first computes a histogram estimate of the graphon, which is then smoothed by some total variation minimization.

Generally, a simple way to derive an estimator for multiple networks consists in computing a graphon estimator on every network and then taking the average of all graphon estimates. However, this approach only makes sense when all individual graphon estimates have a canonical form. Besides considering the average of individual graphon estimates, none of the existing algorithms mentioned so far is appropriate to directly analyze multiple networks without node correspondence, since the joint analysis of several graphs is involved.

**Contributions.** We make two contributions in this paper. First, for any exchangeable graph model satisfying some identifiability condition, we introduce a novel and efficient method for estimating the graphon on multiple networks with different node sets. The proposed algorithm outperforms the approach where one applies a state-of-the-art algorithm to each of the networks and then takes their average. Second, the proposed estimator demonstrates superior performance, especially as the number of graphs tends to infinity. Notably, there is a significant improvement over existing state-of-the-art algorithms, which fix the number of graphs and where the network size tends to infinity. From this perspective, the proposed algorithm is

computationally more efficient compared to the greedy algorithm proposed by Rebafka (2024) on SBMs.

The rest of this work is organized as follows. In Section 2, we establish the framework of the study and discuss the graphon identifiability issue. In Section 3, we introduce and discuss the new estimator. Finally, in Section 4, we present the simulation study to assess its performance and compare to the state of the art.

# 2 Model

We observe a collection of networks $\mathcal{G} = (G^{(1)}, ..., G^{(M)})$, where the $m$-th graph $G^{(m)} = (V^{(m)}, E^{(m)})$ has node set $V^{(m)}$ and edge set $E^{(m)}$. We assume that the graphs are binary, undirected and do not have the same node sets $V^{(m)}$. The numbers of nodes $n_m = |V^{(m)}|$ may not be identical either. We denote the adjacency matrix of graph $G^{(m)}$ by $A^{(m)} \in \{0,1\}^{n_m \times n_m}$.

Let $\mathcal{G}$ be a set of independent and identically distributed exchangeable random graphs generated from some unknown graphon $w$. A **graphon** $w$ is a symmetric measurable function $w : [0,1]^2 \to [0,1]$, where $w(u,v)$ represents the probability of an edge between nodes of the graph. For each $m$, $G^{(m)}$ is generated by the following sampling scheme. First, generate a uniformly distributed latent variable $U_i^{(m)}$ for each node $i \in \{1, ..., n_m\}$, that is,

$$U_1^{(m)}, ..., U_{n_m}^{(m)} \overset{iid}{\sim} \text{Uniform}[0,1].$$

Then, conditionally to these latent variables, the entries of adjacency matrix $A^{(m)}$ are generated from a Bernoulli distribution with parameter given by the graphon $w$. More precisely,

$$A_{ij}^{(m)} | U_i^{(m)}, U_j^{(m)} \overset{ind}{\sim} \text{Bernoulli}(w(U_i^{(m)}, U_j^{(m)})) \quad \text{for } i \leq j.$$

The goal is to estimate the graphon $w$ from the data $\mathcal{G}$. For the estimation problem to be well-posed, we focus on identifiable graphons as defined below.

**Condition 2.1** (**Strict monotonicity of degrees**). *A graphon $w$ has a unique representation if there exists a measure-preserving transformation $\varphi$ such that $w^{can}(u,v) = w(\varphi(u), \varphi(v))$ and*

$$g^{can}(u) = \int_0^1 w^{can}(u,v)dv,$$

*is strictly increasing. The graphon $w^{can}$ is called the* ***canonical representation*** *of $w$.*

In the rest of this work, we mainly focus on graphons satisfying the strict monotonicity condition. For notational simplicity, we denote $w$ as the canonical representation rather than $w^{can}$.

# 3 Joint Graph Sorting algorithm

The joint graph sorting (JGS) algorithm exploits the fact that the canonical graphon is identifiable and that the latent positions $U_i^{(m)}$ may be recovered by sorting the empirical degrees of the graphs. Then, using the estimated latent positions, a histogram estimate of the graphon is easily defined. This approach is used by several authors, such as Chan and Airoldi (2014), but by performing the sorting network by network. Here, we propose an estimation

procedure based on a joint analysis of the networks. Namely, we consider ordering the nodes not networkwisely, but establishing the latent position of each node with respect to the entire collection of networks. This improves the estimation of the latent variables, and consequently the estimation of the graphon.

**Joint sorting stage**: We compute the empirical degrees for all nodes of all the graphs. In order to put them on the same scale, we divide the empirical degrees by the number of nodes in the graph, and we refer to them as the **normalized empirical degrees**. Formally, we compute for $m = 1, ..., M$,

$$d_i^{(m)} = \frac{1}{n_m} \sum_{j=1}^{n_m} A_{i,j}^{(m)}, \text{ for } i = 1, ..., n_m.$$

Then, we consider the set of all nodes $V = \cup_{m=1}^{M} V^{(m)}$ and order them according to their normalized empirical degrees. More precisely, let $d^{(m)} = (d_1^{(m)}, ..., d_{n_m}^{(m)}) \in [0,1]^{n_m}$ be the sequence of normalized degrees of $G^{(m)}$, for $m = 1, ..., M$. Then we relabel the nodes in the following way. The nodes of the first graph $G^{(1)}$ are denoted by $\{1, ..., n_1\}$. Those of the second graph $G^{(2)}$ are denoted by $\{n_1 + 1, ..., n_1 + n_2\}$ and so on. Now, let $d = (d_1, ..., d_n) = (d_1^{(1)}, ..., d_{n_1}^{(1)}, ..., d_1^{(M)}, ..., d_{n_M}^{(M)})$ be the vector representing the normalized degrees of the nodes of all $M$ graphs, where

$$n = \sum_{m=1}^{M} n_m = |V| = \sum_{m=1}^{M} |V^{(m)}|.$$

To recover the canonical graphon, we search a permutation $\hat{\sigma}$ of the nodes $\{1, ..., n\}$ such that $d_{\hat{\sigma}(1)} \leq \cdots \leq d_{\hat{\sigma}(n)}$. Ordering the nodes according to $\widehat{\sigma}$ amounts to estimate the latent positions by

$$\widehat{U}_i = \frac{\widehat{\sigma}^{-1}(i)}{n} - \frac{1}{2n}, \quad i = 1, ..., n. \tag{3.1}$$

Thus, compared to the previous graphon estimate, a finer grid for the values of the latent positions is considered.

Now, to compute a histogram estimate, we divide the interval $[0,1]$ in $k$ regular intervals of length $h = \frac{1}{k}$ and compute the edge frequencies per block. Formally, let $J_l^{(m)} \subset \{1, ..., n_m\}$ be the set of node indices $i_m$ of graph $G_{(m)}$ such that $\widehat{U}_i = \widehat{U}_{i_m}^{(m)} \in I_l = [(l-1)h, lh[$. Then, the edge frequency of block $I_s \times I_t$, $s, t = 1, ..., k$ is giving by

$$\widehat{H}_{s,t} = \frac{\sum_{m=1}^{M} \sum_{(i,j) \in J_s^{(m)} \times J_t^{(m)}} \widehat{A}_{ij}^{(m)}}{\max\left\{\sum_{m=1}^{M} |J_s^{(m)}| \cdot |J_t^{(m)}|, 1\right\}}. \tag{3.2}$$

Finally, the graphon estimate is giving by

$$\widehat{w}(u, v) = \widehat{H}_{s,t}, \ \forall (u, v) \in I_s \times I_t.$$

This estimation procedure is summarized in Algorithm 1. Compared to the methods by Rebafka (2024) and Chabert-Liddell et al. (2023), which fit a SBM to a collection of networks by using cumbersome iterative algorithms, our estimator is extremely fast as it consists of only two steps. It is also scalable to very huge datasets with a large number of networks and/or large network sizes.

**Algorithm 1:** Joint Graph Sorting algorithm

| | |
|---|---|
| **Input** | : A collection of observed adjacency matrices $A^{(1)}, ..., A^{(M)}$ of respective sizes $n_1, ..., n_M$ and the desired number of histogram blocks $k$ |
| **Joint stage** | : Compute the normalized empirical degrees $d_i^{(m)} = \frac{1}{n^{(m)}} \sum_{j=1}^{n^{(m)}} A_{i,j}^{(m)}$, $i = 1, ..., n_m, m = 1, ..., M$; Order all nodes according to their normalized degrees ; Compute the partition $(I_1, ..., I_k)$, and the sets $(J_1^{(m)}, ..., J_k^{(m)})$; |
| **Global estimate:** | For $s, t \in \{1, ..., k\}$, compute $\widehat{H}_{s,t}$ according to Equation (3.2) |
| **Output** | : Graphon estimate in form of matrix $\widehat{H} \in [0, 1]^{k \times k}$. |

# 4 Simulations study

In this section, we illustrate the performance of the proposed estimator for the continuous graphon $w(u, v) = uv$ that satisfies Condition 2.1 on the monotonicity of the degree sequence.

We compare our estimator to the one which is obtained by computing the SAS estimator by Chan and Airoldi (2014) on each network and then taking the mean of the individual estimators. That is, the node sorting is done networkwisely, and not on the level of the entire set of networks as far as our estimator. Graphs were generated based on this graphon, and we use the mean integrated squared error (MISE) to compare our estimator to the SAS estimator of Chan and Airoldi (2014) in two settings. In the first one, the number $M = 20$ of graphs within each collection is fixed and the graph size $n$ increases from 50 to 1000 (Figure 1.b). We remark that our JGS estimator is consistent and outperforms the SAS estimator. In the second setting, we vary the number $M$ of graphs within each collection, while simultaneously keeping the number $n$ of nodes in all graphs constant. More precisely, we generate collections where every graph has $n = 20$ nodes and the number $M$ of networks increases from 17 to 500 (Figure 1.a). Several observations can be done. First, we see that as expected, the MISE of both estimators decreases when $M$ increases. For every number of networks $M$, the proposed estimator largely outperforms the SAS estimator. Second, it is interesting to see that the MISE of both estimators does not vanish, indicating that none of the estimators is consistent. To understand this fact, it is instructive to analyze the estimators $\widehat{U}_i^{(m)}$ of the latent positions. Figure 2 represents the mean squared error (MSE) of the estimates $\widehat{U}_i^{(m)}$ for the two estimation procedures. In Fig 2.b we see that the latent positions are consistently estimated when adding nodes to each network. However, this is not the case in the second asymptotic setting (see Fig 2.a and its zoom). Indeed, adding now networks to the collection does not improve the normalize node degree $d_i^{(m)}$ as they are only computed on the adjacency matrix $A^{(m)}$ which remains fixed.

# 5 Conclusion

We have proposed an approach for nonparametric graphon estimation based on histograms in the multiple network setting without correspondence of the node sets. Estimating the graphon in this context is a challenging task, primarily due to the problem of graphon non-identifiability and the intricacy of network comparison. Our estimator is very fast and has significantly better accuracy than the SAS estimator. This improvement is due to a better exploitation of the data. Nevertheless, our estimator suffers from some limitations in the asymptotic setting where the numbers of networks increases, while the network sizes are bounded. In

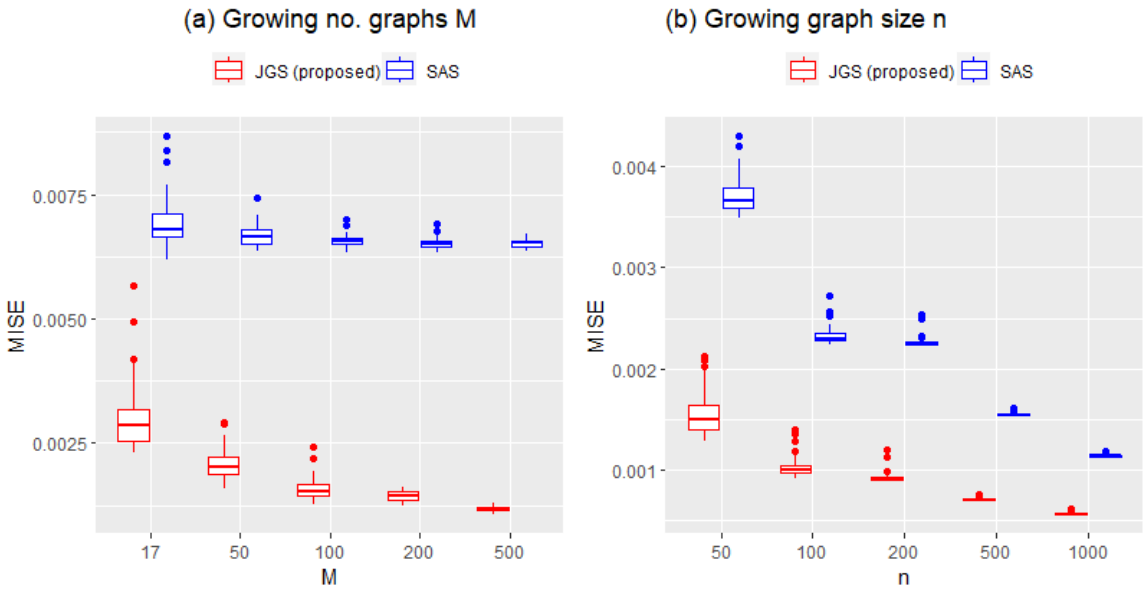



Figure 1: MISE of proposed (JGS) alogrithm vs MISE of SAS algorithm

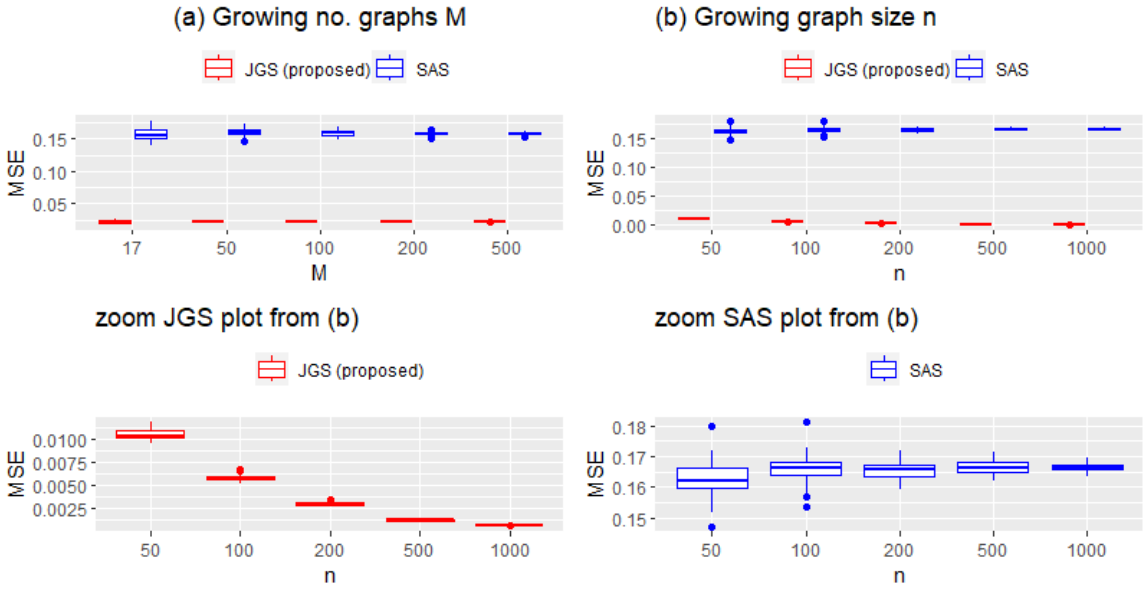



Figure 2: MSE of latent positions: the proposed (JGS) algorithm vs the SAS algorithm

this setting the data do not provide sufficiently much information to consistently estimate the latent node positions.

# References


Airoldi, E. M., Costa, T. B., and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 692–700.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman-girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, 106(50):21068–21073.

Chabert-Liddell, S.-C., Barbillon, P., and Donnet, S. (2023). Learning common structures in a collection of networks: An application to food webs. ArXiv:2206.00560, Mar.

Chan, S. and Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216.

Chatterjee, S. (2012). Matrix estimation by universal singular value thresholding. ArXiv:1212.1247.

Diaconis, P. and Janson, S. (2007). Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni, Series VII*, pages 33–61.

Lloyd, J. R., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pages 1007–1015.

Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96:933–957.

Rebafka, T. (2024). Model-based clustering of multiple networks with a hierarchical algorithm. *Statistics and Computing*, 34:1–16.

# SBM for point set registration

Giulia Marchello[1], Marco Corneli[2,3], Davide Adamo[2,3] & Charles Bouveyron[2]

[1] *Inria PreMeDICaL team, Idesp, Université de Montpellier, France.*

[2] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France.*

[3] *Université Côte d'Azur, Laboratoire CEPAM, Nice, France.*

**Résumé.** L'enregistrement de nuages de points est une tâche fondamentale en vision avec des applications dans la recherche d'images 3D, la segmentation et la reconnaissance de formes. Cet article aborde les défis de l'enregistrement d'ensembles de points, en tenant compte de facteurs tels que les transformations spatiales non rigides, la grande dimensionalité, le bruit et les valeurs aberrantes. L'accent est mis sur les méthodes basées sur des probabilitées, en exploitant le modèle à blocs stochastiques (SBM). L'approche proposée consiste à représenter les nuages de points sous forme de graphes et à introduire une variable latente pour le regroupement. Pour améliorer la parcimonie et l'efficacité computationnelle, le modèle intègre une distribution normale vec excès de zéros, se concentrant uniquement sur les entrées non nulles en dessous d'un seuil spécifié. L'article décrit la distribution conjointe et présente un algorithme d'inférence variationnelle pour l'estimation des paramètres. La méthodologie offre un cadre probabiliste pour un enregistrement robuste d'ensembles de points, démontrant son potentiel dans des scénarios complexes avec des données de grande dimensionalité. En tant que document de travail, les étapes suivantes concerneront des exemples numériques sur des ensembles de données simulés et réels.

**Mots-clés.** Enregistrement de Nuages de Points, Modèle à Blocs Stochastiques, Distribution Normale avec excès de zéros, Inférence Variationnelle, Transformations Spatiales.

**Abstract.** The registration of point clouds is a fundamental task in computer vision with applications in 3D image retrieval, segmentation, and shape recognition. This paper addresses the challenges of point set registration, considering factors such as nonrigid spatial transformations, high dimensionality, noise, and outliers. The focus is on probability-based methods, specifically leveraging the Stochastic Block Model (SBM). The proposed approach involves representing point clouds as graphs and introducing a latent variable for clustering. The proposed approach involves representing point clouds through graphs, introducing a latent variable for clustering. To enhance sparsity and computational efficiency, the model incorporates a Zero-Inflated Normal distribution, focusing solely on non-zero entries below a specified threshold. The paper outlines the joint distribution and presents a variational inference algorithm for parameter estimation. The methodology provides a probabilistic framework for robust point set registration, demonstrating its potential in complex scenarios with high-dimensional data. Being a working paper, the following steps will concern numerical examples on simulated and real dataset.

**Keywords.** Point Cloud Registration, Stochastic Block Model, Zero-Inflated Normal Distribution, Variational Inference, Spatial Transformations.

# 1 Introduction

Registration of point clouds is crucial in various computer vision applications, playing a pivotal role in tasks such as 3D image retrieval, segmentation, and shape recognition. The main goal of point set registration consists in establishing correspondences between two sets of points and determining transformations that maps one point set onto another. Point set registration is a challenging task due to several factors, such as the presence of an unknown nonrigid spatial transformation, the high dimensionality of point sets, potential noise, and the existence of outliers. The transformation considered in point set registration typically falls in two categories: rigid or nonrigid. A rigid transformation is constrained to translation, rotation, and scaling. On the other hand, nonrigid transformations encompass a broader range of alterations, such as stretching and skewing. Figure 1 shows an example of point set registration with a 2D toy example. In this picture, the second image is visibly rotated by 90 degrees and it and contains an additional point denoted by a question mark compared to the first image. This discrepancy adds complexity to the matching task. Extending this concept to more intricate scenarios involving high-dimensional data, noise, outliers, and 3D images, it becomes evident that point sets registration can quickly become a challenging task.

Regarding the pairwise point set registration, depending on the modeling assumptions, several approaches have been proposed in the literature. They can be classified into distance-based methods (Zhang, 1994; Zhou and De la Torre, 2015), filter-based methods (Zhu et al., 2018; Li et al., 2016) and probability-based methods (Myronenko and Song, 2010; Zhou et al., 2014). It has been observed that probability-based methods tend to outperform other approaches. However, it is worth noting that the probability-based methods come at a higher computational cost in contrast to distance-based and filter-based methods (Zhu et al., 2019).

This section aims to formulate a probability-based point set registration method, designed to effectively address rigid transformations in the context of 3D point clouds.

## 1.1 Our contribution

Our approach would be to use a probabilistic method as well. Specifically, we plan on making use of the Stochastic Block Model instead of Gaussian Mixture Models (GMMs) in order to model the points of a source cloud with those of a target cloud. In more detail, we represent a point cloud with $N$ points via a graph with $N$ nodes, thus leading to a (weighted) adjacency $N \times N$ matrix, denoted by $X$. The way the nodes are connected to each other and hence the nature of the graph is of course crucial and several approaches are possible. The idea is indeed to use a an $\epsilon$-neighborhood graph where $\Theta_{ij}$ being either the Euclidean distance between to neighbor nodes or zero, with: $i, j = 1, \ldots, N$:

$$\Theta_{ij} = \begin{cases} d(i,j) & \text{if } d(i,j) < \epsilon, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Also, one might contemplate using a fully connected graph, where $X_{ij}$ denotes the geodesic distance between points $i$ and $j$. However, it is imperative to acknowledge that the compu-

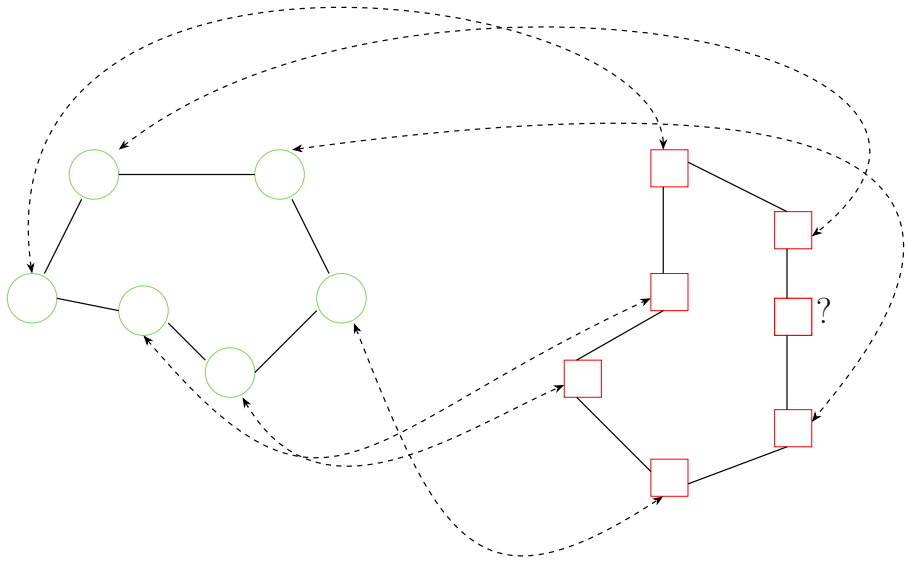



Figure 1: A toy example of the point set registration problem.

tational cost associated with this approach tends to escalate proportionally with the growing number of points.

For simplicity, we focus on two point clouds $\mathcal{X}$ and $\mathcal{Y}$ represented by two matrices of dimension $N \times 3$ and $Q \times 3$, respectively, with $N > Q$. We compute their distance adjacency matrices following Eq.(1), obtaining $\Theta^{\mathcal{X}}$ and $\Theta^{\mathcal{Y}}$, respectively.

Now, let $Z$ be a latent matrix such that: $Z := \{z_{iq}\}_{i \in 1,\dots,N, q \in 1,\dots,Q}$. This matrix represents the clustering of point of $\mathcal{X}$ into $Q$ groups, where point $i$ belongs to cluster $q$ if $z_{iq} = 1$, 0 otherwise. Moreover, the rows of $Z$ are assumed to be independently distributed according to a multinomial distributions:

$$p(Z|\alpha) = \prod_{i=1}^{N} \prod_{q=1}^{Q} \alpha_q^{z_{iq}},$$

where $\alpha_q = \mathbb{P}\{z_{iq} = 1\}$ and $\sum_{q=1}^{Q} \alpha_q = 1$.

We further assume that, conditionally to $Z$, $\Theta^{\mathcal{X}}$ follows a Zero-Inflated Normal distribution, such that:

$$\Theta^{\mathcal{X}}_{ij} | Z_i Z_j \sim \mathcal{ZIN}(\Theta^{\mathcal{Y}}_{Z_i Z_j}; \sigma^2_{Z_i Z_j}). \tag{2}$$

Being a mixture between a chosen distribution and a Dirac mass at zero, the Zero-Inflated

distribution is used to account for a high sparsity in the data and can be formally written as:

$$\begin{cases} \Theta^{\mathcal{X}}_{ij}|Z_i, Z_j \sim \delta_0(\Theta^{\mathcal{X}}_{ij}) & \text{with probability } \pi_{Z_iZ_j} \\ \Theta^{\mathcal{X}}_{ij}|Z_i, Z_j \sim N(\Theta^{\mathcal{X}}_{ij}; \Theta^{\mathcal{Y}}_{Z_iZ_j}; \sigma^2_{Z_iZ_j}) & \text{with probability } 1 - \pi_{Z_iZ_j}. \end{cases} \tag{3}$$

where $\delta_0(\cdot)$ is the Dirac function in 0 and $\Theta^{\mathcal{Y}}_{Z_iZ_j}, \sigma^2_{Z_iZ_j}$ are the block-dependent parameters of the Normal distribution.

Then, to model the sparsity, we rewrite Eq. (3) by introducing a hidden random matrix, $A \in \{0,1\}^{N\times N}$, where:

$$A_{ij}|Z_iZ_j \sim \mathcal{B}(\pi_{Z_iZ_j}),$$

with $\mathcal{B}(\pi)$ denoting the Bernoulli probability mass function of parameter $\pi$ and such that

$$\begin{aligned} A_{ij} = 1 &\Rightarrow \Theta^{\mathcal{X}}_{ij}|Z_i, Z_j \sim \delta_0(\Theta^{\mathcal{X}}_{ij}) \\ A_{ij} = 0 &\Rightarrow \Theta^{\mathcal{X}}_{ij}|Z_i, Z_j \sim N(\Theta^{\mathcal{X}}_{ij}; \Theta^{\mathcal{Y}}_{Z_iZ_j}; \sigma^2_{Z_iZ_j}). \end{aligned} \tag{4}$$

# 2 The joint distribution

The model described so far has a set of parameters denoted by $\theta := (\Theta^{\mathcal{Y}}, \Sigma, \alpha, \pi)$, with $\Theta^{\mathcal{Y}} =: \{\Theta^{\mathcal{Y}}_{ql}\}_{q,l\leq Q}$ and $\Sigma := \{\sigma^2_{ql}\}_{q,l}$ and a set of latent variables: $\{Z,\ A\}$.
As a first move, we can compute the likelihood of the complete data:

$$p(\Theta^{\mathcal{X}}, A, Z|\theta) = p(\Theta^{\mathcal{X}}|A, Z, \Theta^{\mathcal{Y}}, \Sigma)p(A|Z, \pi)p(Z|\alpha), \tag{5}$$

where:

$$p(\Theta^{\mathcal{X}}|A, Z, \Theta^{\mathcal{Y}}, \Sigma) = \prod_{j>i}^{N} \mathbf{1}^{A_{ij}}_{\{\Theta^{\mathcal{X}}_{ij}=0\}} \left\{ \left( \mathcal{N}(\Theta^{\mathcal{X}}_{ij}; \Theta^{\mathcal{Y}}_{Z_iZ_j}; \sigma^2_{Z_iZ_j}) \right)^{(1-A_{ij})} \right\},$$

$$p(A|Z, \pi) = \prod_{j>i}^{N} \pi^{A_{ij}}_{Z_iZ_j} \left(1 - \pi_{Z_iZ_j}\right)\Big)^{(1-A_{ij})},$$

$$p(Z|\alpha) = \prod_{i=1}^{N} \prod_{q=1}^{Q} \alpha_q^{Z_{iq}},$$

where the shorthand notation $\sum_{j>i}^{N} := \sum_{i=1}^{N} \sum_{j>i}^{N}$ was adopted. By looking at the first of the above equations, it is immediately clear that if $A_{ij} = 1$ and $\Theta^{\mathcal{X}}_{ij} \neq 0$ the whole likelihood goes to zero. On the contrary, we assume that $\mathbf{1}^{A_{ij}}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}$ takes value one when both $A_{ij}$ and the indicator function are zero. So, in order to have a positive likelihood, we must assume that we never observe $A_{ij} = 1$ and $\Theta^{\mathcal{X}}_{ij} \neq 0$ and, as a consequence, we could replace the indicator function with 1. However we let it in place since it will provide us with an interesting

intuition, later. Hence, we can write the log-likelihood of the complete data as follows:

$$\begin{aligned}\log p(\Theta^{\mathcal{X}}, Z, A|\theta) = \frac{1}{2}\sum_{j\neq i}^{N}\sum_{q,\ell}^{Q}\Big[A_{ij}Z_{iq}Z_{j\ell}\log\left(\pi_{ql}\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}\right)+ \\ + (1-A_{ij})Z_{iq}Z_{j\ell}\log\left((1-\pi_{ql})\mathcal{N}(\Theta^{\mathcal{X}}_{ij};\Theta^{\mathcal{Y}}_{ql},\sigma^2_{ql})\right)\Big]+ \\ + \sum_{i=1}^{N}\sum_{q=1}^{Q}Z_{iq}\log\alpha_q.\end{aligned} \tag{6}$$

# 3 The inference

## 3.1 Variational assumptions

Since we cannot compute the posterior distribution, $p(A, Z|\Theta^{\mathcal{X}}, \theta)$, we rely on a variational procedure which optimizes a lower bound of the likelihood. Let us thus introduce a variational distribution $q(.)$ in order to decompose the likelihood as follows:

$$\log p(\Theta^{\mathcal{X}}|\theta) = \mathcal{L}(q,\theta) + KL(q \parallel p_{A,Z}),$$

where $\mathcal{L}$ denotes a lower bound defined as

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_A\sum_Z q(A,Z)\log\frac{p(\Theta^{\mathcal{X}}, A, Z|\theta)}{q(A,Z)} \\ &= E_{q(A,Z)}\Big[\log\frac{p(\Theta^{\mathcal{X}}, A, Z|\theta)}{q(A,Z)}\Big] \\ &= E_{q(A,Z)}[\log(p(\Theta^{\mathcal{X}}, A, Z|\theta)] - E_{q(A,Z)}[\log(q(A,Z))],\end{aligned} \tag{7}$$

and KL indicates the Kullaback-Liebler divergence between the true and the approximate posterior $p_{A,Z} := p(A, Z|\Theta^{\mathcal{X}}, \theta)$:

$$KL(q \parallel p_{A,Z}) = -\sum_A\sum_Z q(A,Z)\log\frac{p(A,Z|\Theta^{\mathcal{X}},\theta)}{q(A,Z)}.$$

Now, the objective is to find a distribution $q(\cdot)$ that maximizes the lower bound $\mathcal{L}(q,\theta)$. In order to allow the optimization of $\mathcal{L}(q,\theta)$, we further assume that $q(A,Z)$ factorizes as follows:

$$\begin{aligned}q(A,Z) = q(A)q(Z) &= \prod_{j\neq i}^{N} q(A_{ij})\prod_{i=1}^{N} q(Z_i) \\ &= \prod_{j\neq i}^{N}\delta_{ij}^{A_{ij}}(1-\delta_{ij})^{(1-A_{ij})}\prod_{i=1}^{N}\prod_{q}^{Q}\tau_{iq}^{Z_{iq}},\end{aligned}$$

Where $\delta_{ij}$ and $\tau_{iq}$ indicate the variational parameters of $A_{ij}$ and $\tau_{iq}$, respectively (see Section 3.2 for details).

## 3.2 VE-Step

The VE-step of the VEM algorithm aims at maximizing the lower bound in Eq. (7) with respect to the variational distribution $q(\cdot)$ while keeping $\theta$ fixed. Following **?**, we derive the update equations for the factors $q(A)$ and $q(Z)$, such that the log of the optimized factors are given by:

$$\log q^*(A) = E_{q(Z)}[\log p(\Theta^{\mathcal{X}}, A, Z|\theta)], \tag{8}$$

$$\log q^*(Z) = E_{q(A)}[\log p(\Theta^{\mathcal{X}}, A, Z|\theta)], \tag{9}$$

### 3.2.1 Optimization of the factor q(A)

Let us consider the derivation of the update equation for the factor $q(A)$. The sequential update for the factor $q(A)$ can be computed through the logarithm of the optimized factor, where all the terms that do not depend on $A$ are absorbed in a constant.

**Proposition 1.** *Denoting by $\delta_{ij}$ the variational success probability for $A_{ij}$, the optimal update of $q(A)$ is:*

$$\delta_{ij} = \frac{\exp(R_{ij})}{1 + \exp(R_{ij})}, \tag{10}$$

*with:*

$$R_{ij} := \log \mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}} + \frac{1}{2} \sum_{q,\ell}^{Q} \tau_{iq}\tau_{j\ell} \left( \log\left(\frac{\pi_{q\ell}}{1-\pi_{q\ell}}\right) - \log \mathcal{N}(\Theta^{\mathcal{X}}_{ij}; \Theta^{\mathcal{Y}}_{ql}; \sigma^2_{ql}) \right). \tag{11}$$

*Proof.*

$$\begin{aligned}
\log q^*(A) &= E_{q(Z)}[\log p(\Theta^{\mathcal{X}}, A, Z|\theta)] \\
&= \frac{1}{2} \sum_{j\neq i}^{N} \sum_{q,\ell}^{Q} \Big[ A_{ij}\tau_{iq}\tau_{j\ell} \log\left(\pi_{q\ell}\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}\right) + \\
&+ (1 - A_{ij})\tau_{iq}\tau_{j\ell} \log\left((1-\pi_{q\ell})\mathcal{N}(\Theta^{\mathcal{X}}_{ij}; \Theta^{\mathcal{Y}}_{q\ell}; \sigma^2_{q\ell})\right) \Big] + C \\
&= \frac{1}{2} \sum_{j\neq i}^{N} \sum_{q,\ell}^{Q} \Big[ A_{ij}\tau_{iq}\tau_{j\ell} \log\left(\pi_{q\ell}\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}\right) + \\
&- A_{ij}\tau_{iq}\tau_{j\ell} \log\left((1-\pi_{q\ell})\mathcal{N}(\Theta^{\mathcal{X}}_{ij}; \Theta^{\mathcal{Y}}_{q\ell}; \sigma^2_{q\ell})\right) \Big] + C \\
&= \frac{1}{2} \sum_{j\neq i}^{N} A_{ij} \left[ \sum_{q,\ell}^{Q} \tau_{iq}\tau_{j\ell} \Big( \log\left(\frac{\pi_{q\ell}}{1-\pi_{q\ell}}\right) + \log(\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}) + \right. \\
&\left. - \log(\mathcal{N}(\Theta^{\mathcal{X}}_{ij}; \Theta^{\mathcal{Y}}_{q\ell}; \sigma^2_{q\ell})) \Big) \right] + C.
\end{aligned} \tag{12}$$

□

We can then recognize the functional form of a Bernoulli distribution:

$$\begin{aligned}\log q^*(A) &= \sum_{j\neq i}^{N} A_{ij}\log\delta_{ij} + (1-A_{ij})\log(1-\delta_{ij}),\\ &= \sum_{j\neq i}^{N} A_{ij}\log\left(\frac{\delta_{ij}}{1-\delta_{ij}}\right) + C,\end{aligned} \tag{13}$$

where $\delta_{ij}$ is

$$\delta_{ij} = \frac{\exp(R_{ij})}{1+\exp(R_{ij})}, \tag{14}$$

with $R_{ij}$ defined in Eq. (11).

Note that, although everything in the above equations is well defined in force of the assumptions we made (in particular one never has $\log(0)$ unless $A_{ij}=0$) formally, when $\Theta^{\mathcal{X}}_{ij}\neq 0$, $R_{ij}=-\infty$ and $\delta_{ij}=0$, which makes sense: non-null distances in $\Theta^{\mathcal{X}}_{ij}$ come from a Normal distribution distribution with probability one (see Eq. 4).

### 3.2.2 Optimization of the factor q(Z)

Let us consider the derivation of the update equation for the factor $q(Z)$. The sequential update for the factor $q(Z)$ can be computed through the log of the optimized factor.

**Proposition 2.** *Denoting by $\tau_{iq}$ the variational success probability for $Z_{iq}$ the optimal update of $q(Z)$ is:*

$$\tau_{iq} = \frac{r_{iq}}{\sum_{\ell=1}^{Q} r_{i\ell}}, \tag{15}$$

*with:*

$$r_{iq} \propto \alpha_q \exp\left(\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\ell=1}^{Q}\tau_{j\ell}\left[\delta_{ij}\log\left(\pi_{q\ell}\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}\right) + (1-\delta_{ij})\log\left((1-\pi_{q\ell})\mathcal{N}(\Theta^{\mathcal{X}}_{ij};\Theta^{\mathcal{Y}}_{q\ell},\sigma^2_{q\ell})\right)\right]\right) \tag{16}$$

*Proof.* First, we know that

$$\log q^*(Z_i) = E_{A,Z_{/i}\sim q}[\log p(\Theta^{\mathcal{X}}, A, Z|\theta))]$$

where the subscript $A, Z_{/i}\sim q$ means that the expectation is taken with respect to all the latent variables following the variational distribution *except* $Z_i$. Thence we can isolate the terms in $\log p(\Theta^{\mathcal{X}}, A, Z|\theta))$ depending on $Z_i$ and take the expectation

$$\begin{aligned}\log q^*(Z_i) = &\sum_{j\neq i}^{N}\sum_{q,\ell}^{Q} Z_{iq}\left(\delta_{ij}\tau_{j\ell}\log\left(\pi_{q\ell}\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}\right) + (1-\delta_{ij})\tau_{j\ell}\log\left((1-\pi_{q\ell})\mathcal{N}(\Theta^{\mathcal{X}}_{ij};\Theta^{\mathcal{Y}}_{q\ell},\sigma^2_{q\ell})\right)\right)\\ &+\sum_{q=1}^{Q} Z_{iq}\log\alpha_q + C\end{aligned}$$

where $C$ regroups all the terms not depending on $Z_i$ and we stress that here $\sum_{j\neq i}^{N}$ is $\sum_{\substack{j=1\\ j\neq i}}^{N}$, with $i$ being fixed. The above equation factorizes as follows

$$\log q^*(Z_i) = \sum_{q=1}^{Q} Z_{iq} \log r_{iq} + C \tag{17}$$

where

$$\log r_{iq} := \sum_{j\neq i}^{N}\sum_{\ell=1}^{Q} \tau_{j\ell}\left(\delta_{ij}\log\left(\pi_{q\ell}\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}\right) + (1-\delta_{ij})\log\left((1-\pi_{q\ell})\mathcal{N}(\Theta^{\mathcal{X}}_{ij};\Theta^{\mathcal{Y}}_{q\ell},\sigma^2_{q\ell})\right)\right) + \log\alpha_q.$$

We recognize the log-likelihood of a multinomial distribution in Eq. (17) and taking the exponential on both sides of the above equation the proposition is proven. □

## 3.3 M-Step

In order to obtain the updating of the parameter set $\theta$, the objective of the M-Step is the maximization of the lower bound $L(q;\theta)$ with respect to $\alpha$ and $\pi$. It is worth noticing that the parameter $\Theta^{\mathcal{Y}}_{q\ell}$ is observed (see Section 1.1) and that at this stage we fix the variance of the Normal distribution, $\Sigma$.
The final expression of the variational lower bound can be obtained by developing Eq. (7) as follows:

$$\begin{aligned}\mathcal{L}(q,\theta) &= E_{q(A,Z)}\Big[\log p(X,A,Z|\theta)\Big] - E_{q(A,Z)}\Big[\log q(A,Z)\Big]\\ &= \frac{1}{2}\sum_{j\neq i}^{N}\sum_{q,\ell}^{Q}\left[\delta_{ij}\tau_{iq}\tau_{j\ell}\log(\pi_{q\ell}\mathbf{1}_{\{\Theta^{\mathcal{X}}_{ij}=0\}}) + (1-\delta_{ij})\tau_{iq}\tau_{j\ell}\log(1-\pi_{q\ell})\mathcal{N}(\Theta^{\mathcal{X}}_{ij};\Theta^{\mathcal{Y}}_{q\ell},\sigma^2_{q\ell})\right] +\\ &+\sum_{i\neq 1}^{N}\sum_{q=1}^{Q}\tau_{iq}\log\alpha_q - \sum_{i\neq 1}^{N}\sum_{q=1}^{Q}\tau_{iq}\log\tau_{iq} - \sum_{j\neq i}^{N}\Big[\delta_{ij}\log(\delta_{ij}) + (1-\delta_{ij})\log(1-\delta_{ij})\Big]\end{aligned} \tag{18}$$

### 3.3.1 Update of $\pi$:

Here our goal is to derive the update of the block-dependent sparsity parameter, $\pi_{q\ell}$. The variational distribution $q(A,Z)$ is kept fixed, while the lower bound in Eq.(18) is maximized with respect to $\pi_{q\ell}$, to obtain its update,$\hat{\pi}_{q\ell}$.

**Proposition 3.** *The updating formula of $\pi$ is obtained by maximizing $L(q;\theta)$ with respect to the parameter. Following algebraic manipulations, it can be expressed as:*

$$\hat{\pi}_{q\ell} := \frac{\sum_{j\neq i}^{N}\delta_{ij}\tau_{iq}\tau_{j\ell}}{\sum_{j\neq i}^{N}\left[\tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta^{\mathcal{X}}_{ij};\Theta^{\mathcal{Y}}_{q\ell},\sigma^2_{q\ell}) + \delta_{ij}\tau_{iq}\tau_{j\ell} - \delta_{ij}\tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta^{\mathcal{X}}_{ij};\Theta^{\mathcal{Y}}_{q\ell},\sigma^2_{q\ell})\right]} \tag{19}$$

*Proof.*

$$
\begin{aligned}
\frac{\partial \mathcal{L}(q,\theta)}{\partial \pi_{q\ell}} &= \frac{1}{2}\sum_{j\neq i}^{N}\left[\frac{\delta_{ij}\tau_{iq}\tau_{j\ell}}{\pi_{q\ell}} - \frac{(1-\delta_{ij})\tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta_{ij}^{\mathcal{X}};\Theta_{q\ell}^{\mathcal{Y}},\sigma_{q\ell}^2)}{1-\pi_{q\ell}}\right] = 0 \\
&= \sum_{j\neq i}^{N}\Bigg[\delta_{ij}\tau_{iq}\tau_{j\ell} - \delta_{ij}\tau_{iq}\tau_{j\ell}\pi_{q\ell} - \pi_{q\ell}\tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta_{ij}^{\mathcal{X}};\Theta_{q\ell}^{\mathcal{Y}},\sigma_{q\ell}^2) + \\
&+ \delta_{ij}\tau_{iq}\tau_{j\ell}\pi_{q\ell}\mathcal{N}(\Theta_{ij}^{\mathcal{X}};\Theta_{q\ell}^{\mathcal{Y}},\sigma_{q\ell}^2)\Bigg] = 0 \\
&= \sum_{j\neq i}^{N}\pi_{q\ell}\left[-\delta_{ij}\tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta_{ij}^{\mathcal{X}};\Theta_{q\ell}^{\mathcal{Y}},\sigma_{q\ell}^2) + \tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta_{ij}^{\mathcal{X}};\Theta_{q\ell}^{\mathcal{Y}},\sigma_{q\ell}^2) + \delta_{ij}\tau_{iq}\tau_{j\ell}\right] = \sum_{j\neq i}^{N}\delta_{ij}\tau_{iq}\tau_{j\ell} \\
\hat{\pi}_{q\ell} &:= \frac{\sum_{j\neq i}^{N}\delta_{ij}\tau_{iq}\tau_{j\ell}}{\sum_{j\neq i}^{N}\left[\tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta_{ij}^{\mathcal{X}};\Theta_{q\ell}^{\mathcal{Y}},\sigma_{q\ell}^2) + \delta_{ij}\tau_{iq}\tau_{j\ell} - \delta_{ij}\tau_{iq}\tau_{j\ell}\mathcal{N}(\Theta_{ij}^{\mathcal{X}};\Theta_{q\ell}^{\mathcal{Y}},\sigma_{q\ell}^2)\right]}
\end{aligned}
\tag{20}
$$

□

### 3.3.2 Update of $\alpha$:

Here our goal is to derive the update of the mixing parameter, $\alpha_q$. While maintaining the variational distribution $q(A, Z)$ constant, we aim to maximize the lower bound presented in Eq.(18) concerning $\alpha_{q\ell}$. Consequently, the optimal update is obtained straightforwardly and can be expressed as:

$$\hat{\alpha}_q = \frac{1}{N}\sum_{i=1}^{N}\tau_{iq} \tag{21}$$

# References


L. Li, M. Yang, C. Wang, and B. Wang. Cubature kalman filter based point set registration for slam. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 847–852. IEEE, 2016.

A. Myronenko and X. Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.

Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.

F. Zhou and F. De la Torre. Factorized graph matching. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1774–1789, 2015.

Z. Zhou, J. Zheng, Y. Dai, Z. Zhou, and S. Chen. Robust non-rigid point set registration using student's-t mixture model. *PloS one*, 9(3):e91381, 2014.

H. Zhu, B. Guo, K. Zou, Y. Li, K.-V. Yuen, L. Mihaylova, and H. Leung. A review of point set registration: From pairwise registration to groupwise registration. *Sensors*, 19(5):1191, 2019.

X. Zhu, M. Ding, T. Huang, X. Jin, and X. Zhang. Pcanet-based structural representation for nonrigid multimodal medical image registration. *Sensors*, 18(5):1477, 2018.

# Session groupe Enseignement de la Statistique

# DIDASTAT$_{EXPL}$ : UN PROJET DE RECHERCHE SUR LES PRATIQUES ENSEIGNANTES EN STATISTIQUE DANS LA FORMATION PROFESSIONNELLE DES FUTURS STATISTICIENS

Charlotte Derouet[1], Camille Doukhan[1], Eric Flavier[1], Hussein Sabra[2], Antoine Rolland[3] & Sonia Yvain-Prébiski[4]

[1] *LISEC, UR 2310, Université de Strasbourg, UL, UHA, France, charlotte.derouet@unistra.fr*
[2] *CRESTIC, Université de Reims*
[3] *ERIC, Université Lyon 2, France, antoine.rolland@univ-lyon2.fr*
[4] *S2HEP-INSPE, Université Lyon 1, France*

**Résumé.** Dans cette communication nous présenterons un projet de recherche actuellement en cours qui s'inscrit dans le champ de la didactique des mathématiques et plus largement de l'éducation, portant sur les pratiques enseignantes en statistique. Cette étude exploratoire vise à décrire les pratiques d'une enseignante en statistique dans une formation professionnalisante pour futurs chargés d'étude statistique : le Bachelor Universitaire de Technologie Science des Données. L'enseignement étudié porte sur les techniques de sondage. Nous présenterons le contexte institutionnel du terrain de notre étude, puis nous exposerons nos objectifs de recherche et la méthodologie générale de ce projet pluridisciplinaire (didactique des mathématiques, science de l'éducation et de la formation, statistique).

**Mots-clés.** Pratiques enseignantes, formation des statisticiens, didactique de la statistique, enseignement supérieur.

**Abstract.** In this paper we will present a research project currently underway in the field of mathematics education focusing on teaching practices in statistics. The aim of this exploratory study is to describe the practices of a statistics teacher in a professional training course for future statistical researchers: the *Bachelor Universitaire de Technologie Science des Données*. The teaching studied concerns survey techniques. We will present the institutional context of our study area, then outline our research objectives and the general methodology of this multidisciplinary project (mathematics education, education and training science, statistics).

**Keywords.** teaching practices, statisticians training, statistics education, tertiary education.

## 1 Contexte du projet DidaStat$_{Expl}$

Le développement des métiers liés à la data (data scientist, data analyst, etc.) depuis plusieurs années s'accompagne d'une augmentation des formations à visée professionnelle dans ce domaine : formations courtes ou longues, diplômantes ou non, de niveau Licence ou Master. Ces formations ont pour la plupart une coloration statistique. La question de l'enseignement de la statistique à ces futurs professionnels se pose. Jusqu'à maintenant, les recherches en éducation portant sur l'enseignement et l'apprentissage de la statistique se sont essentiellement intéressées aux étudiants provenant de filières non-spécialistes telles que la psychologie, les sciences de l'éducation, l'économie... (Batanero, 2004 ; Hahn, 2015 ; Quéré, 2022 ; Schwab-

McCoy, 2019). Les recherches spécifiques au public de futurs statisticiens sont encore rares (Biehler et al., 2022, Rolland, 2020). Le projet que nous menons actuellement cherche justement à apporter des éléments de réponses à la question de l'enseignement de la statistique pour ces futurs professionnels.

Le projet de recherche $\text{DidaStat}_{\text{Expl}}$ a pour objectif général d'étudier les pratiques enseignantes en statistique dans un contexte particulier de formation professionnalisante, nous reviendrons sur ce terme dans la partie suivante. Nous nous concentrons dans cette étude sur le Bachelor Universitaire de Technologie (BUT), spécialité Science des Données (SD). Ce projet interdisciplinaire regroupe des chercheurs en didactique des mathématiques, en science de l'éducation et de la formation et en statistique.

L'objectif de cette communication est de présenter le contexte institutionnel particulier de notre terrain d'étude, les objectifs de notre recherche ainsi que la méthodologie générale adoptée pour atteindre notre objectif initial.

# 2 Un contexte institutionnel particulier : le BUT Sciences des données

Le Bachelor Universitaire de Technologie (BUT), spécialité Science des Données (SD), est une formation professionnalisante pour les métiers de chargés d'étude statistique (data analyst). Il s'agit d'une nouvelle formation en trois ans, depuis la rentrée 2022, issue de la réforme des Instituts Universitaires de Technologie (IUT). Cette formation permet l'insertion professionnelle immédiate des diplômés autant que leur poursuite d'étude éventuelle en master, c'est pourquoi nous la qualifions de formation professionnalisante. Enfin, cette dernière possède un cadrage national, commun à l'ensemble des BUT spécialité SD de France.

Dans ce contexte institutionnel, nous cherchons à étudier les pratiques d'enseignement de la théorie des sondages en deuxième année au travers d'une étude de cas (un seul terrain d'observation). Il s'agit d'une première étude exploratoire.

Une analyse curriculaire nous a permis de dégager des éléments importants constitutifs de cette formation particulière. Le BUT (toutes spécialités confondues) possède un curriculum basé sur l'approche par compétences. Cette spécificité de la formation a un impact fort sur les enseignements proposés dans ces formations. Chaque parcours vise l'acquisition de quatre à six compétences sur les trois années. Dans notre cas, pour le parcours Science des Données : exploration et modélisation statistique, les quatre compétences ciblées dans la formation sont :

- Modéliser les données dans un cadre statistique
- Traiter les données à des fins décisionnelles
- Analyser statistiquement les données
- Valoriser une production dans un contexte professionnel

Afin d'acquérir ces compétences, les modules d'enseignement du BUT SD sont structurés autour de *Ressources*, qui correspondent aux enseignements qui permettent « l'acquisition des connaissances et méthodes fondamentales » (programme, Arrêté du 15-4-2022) (sous forme de cours, TD et/ou TP) et des *Situations d'Apprentissage et d'Évaluation* (*SAÉ*), qui correspondent à des mises en situation professionnelle au cours desquelles l'étudiant développe la compétence » (programme, Arrêté du 15-4-2022). Les *SAÉ* sont décrites par Bolou-Chiaravalli et al. (2022) comme des « situations qui ont une visée intégrative confrontant l'étudiant à des activités s'approchant de celles rencontrées par les professionnels sur le terrain », avec en plus

une visée d'évaluation certificative.

La *SAÉ* étudiée dans notre projet est en deuxième année de BUT au semestre 3. Elle s'intitule « Recueil et analyse de données par échantillonnage ou plan d'expérience » (SAE 3.EMS.01) et a pour objectifs de :

- Approfondir la notion d'enquête et de sondage dans un cadre plus général
- Faire comprendre à l'étudiant la différence qu'implique le tirage sans remise, situation la plus classique dans le cadre d'un sondage
- Amener l'étudiant à mener une réflexion sur la mise en place d'un plan d'expérience.

Notre analyse curriculaire nous a amené à considérer la *SAÉ* comme une interface entre les compétences à développer dans la formation, les *Ressources* (cours/TD/TP délivrés dans le BUT) associées et le futur environnement professionnel des étudiants (Derouet et al., soumis). Nous faisons donc l'hypothèse que la *SAÉ* joue un rôle central dans les pratiques enseignantes en statistique dans en BUT SD et avons décidé d'explorer les pratiques d'une enseignante sous cet angle. Nous cherchons à décrire les pratiques de cette enseignante, par le biais des questions suivantes :

- Quels liens sont faits entre la *SAÉ* et les différentes *Ressources* ? comment ? pourquoi ?
- Quels liens sont faits entre la *SAÉ* et les compétences à développer par les étudiants ? comment ? pourquoi ?
- Quels liens sont faits entre la *SAÉ* et les futures pratiques professionnelles des étudiants ? comment ? pourquoi ?

# 3 Présentation générale des tâches et de la méthodologie du projet

Pour aborder ces questions, notre projet s'articule autour de quatre tâches :

- Tâche 1 : Analyse du savoir statistique,
- Tâche 2 : Analyse des pratiques des professionnels des sondages,
- Tâche 3 : Analyse des pratiques enseignantes et interactions avec les supports d'enseignement,
- Tâche 4 : Analyse de l'activité enseignante *in situ.*

L'objectif de la tâche 1 est de caractériser les savoirs statistiques en jeu au travers d'une analyse épistémologique des savoirs associés à la théorie des sondages. Pour cela nous menons une analyse des livres de référence (reconnus comme tels par les enseignants).

La tâche 2 cherche à identifier des éléments invariants dans les pratiques des professionnelles des sondages, afin de mieux caractériser l'activité statistique (en sondage) et d'enrichir nos analyses didactiques des pratiques de l'enseignante dans un second temps. Pour cela, une étude épistémologique des pratiques professionnelles contemporaines des professionnels des sondages, en appui sur une étude épistémologique de l'activité statistique, est conduite. Cette étude prend appui sur des entretiens de plusieurs professionnels. Elle permet d'avoir des éléments pour analyser les observations de classe.

La tâche 3 vise à questionner les pratiques en fonction du contexte, décrire les supports d'enseignement produits, utilisés et mis à disposition par l'enseignante et accéder aux pratiques effectives de l'enseignante et aux facteurs qui déterminent ses pratiques. Pour cette tâche, nous

avons recueilli plusieurs types de données : entretiens avec l'enseignante en amont et en aval de l'enseignement étudié, supports de cours, vidéos des séances de classe (de la *Ressource* et de la *SAÉ*).

Enfin au travers de la tâche 4 nous cherchons à décrire l'activité de l'enseignante telle que vécue par elle, et à accéder au point de vue singulier de l'activité tout en identifiant sa part non visible. Dans cet objectif, nous avons réalisé puis analysé deux entretiens d'auto-confrontation consécutifs à des séances de classe filmées.

Ces quatre tâches complémentaires permettent d'apporter des éléments de réponse quant à l'impact, sur les pratiques enseignantes en statistique, de la structure du BUT SD et des contraintes institutionnelles associées. Plus précisément, nous cherchons ici à identifier, dans les pratiques enseignantes, les liens entre la *SAÉ* d'une part et les *Ressources* associées, les compétences visées dans la formation et l'environnement professionnel d'autre part. Différents cadres théoriques issus de la didactique des mathématiques et des sciences de l'éducation, sont également mobilisés pour reformuler les questions et analyser nos données.

**Remerciements**

Ce projet de recherche est financé par le GIS Education & Formation du Grand-Est (France).

**Bibliographie**

Arrêté du 15-4-2022 relatif aux dispositions générales des programmes nationaux de la licence professionnelle « BUT » et aux programmes nationaux de chacune des spécialités de licence professionnelle « BUT » (JO du 23-4-2022). https://www.enseignementsup-recherche.gouv.fr/fr/bo/22/Special4/ESRS2211617A.htm

Batanero, C. (2004). Statistics Education as a field for research and practice, Regular lecture, ICME10.

Biehler, R., De Veaux, R., Engel, J., Kazak, S. and Frischemeier D. (2022). Special issue: research on data science education. *Statistics Education Research Journal*, *21*(2).

Bolou-Chiaravalli, C., Bournel-Bosson, M., & Lasne, A. (2022). Professionnalisation et processus réflexif dans le nouveau diplôme proposé par les IUT : le Bachelor Universitaire de Technologie. *Communication & Professionnalisation, 13 : Ressources pédagogiques et professionnalisation dans les formations à la communication*, 122-143.

Derouet, C., Doukhan, C., & Sabra, H. (soumis). Analysing statistical teaching practices in a specific institutional context. *INDRUM 2024*.

Hahn, C. & Stoltz, G. (2013). Savoir académique, savoirs pratiques : tensions et recherche d'équilibre. *Statistique et Enseignement*, *4*(2), 19–52.

Hahn, C. (2015). La recherche internationale en éducation statistique : état des lieux et questions vives. *Statistique et Enseignement*, *6*(2), 25-39.

Quéré, P.-V. (2022). Bridging the mathematics gap between the engineering classroom and the workplace. *International Journal of Mathematical Education in Science and Technology, 53*(5), 1190-1212.

Rolland, A (2020). 2009-2019 : 10 ans de didactique de la statistique en France. *Statistique et Société, 8*(1), 55-71.

Rolland, A. (2023). Une approche de transposition didactique pour l'enseignement universitaire du modèle de régression linéaire en statistique. *Recherches en Didactique des Mathématiques*, *43*(2), 171–198.

Schwab-McCoy, A. (2019). The State of Statistics Education Research in Client Disciplines: Themes and Trends Across the University. *Journal of Statistics Education, 27*(3), 253-264.

# PLUS DE LITTÉRATIE STATISTIQUE = PLUS DE CITOYENNETÉ ACTIVE : DONNÉES ET MÉTHODES

Simona Cafieri

*Istat, Italie, cafieri@istat.it*

**Résumé**

Au niveau international, l'importance de la statistique publique est désormais reconnue : elle est indispensable pour que chaque citoyen puisse connaître et reconnaître de manière critique le contexte dans lequel il vit et être capable de prendre des décisions, même en situation d'incertitude, sur la base de données concrètes et partagées.

Les instituts nationaux de statistique, animés par la conviction qu'une société où les décisions sont prises sur la base de données objectives est une société où règne la démocratie, se sont efforcés de rendre l'accès aux données officielles gratuit et facilement accessible à leurs citoyens. Une réponse immédiate donc, mais non sans écueils car le libre accès à l'information statistique est un moyen d'assurer le progrès d'une société démocratique, mais il perd de son efficacité si les citoyens ne sont pas en mesure de lire, d'interpréter et d'analyser les données. Donc.il s'agissait d'un nouveau défi:non seulement diffuser les données, mais les promouvoir, apprendre à les connaître, les faire comprendre. pour qu'elles soient utiles à la communauté, un bien qui a un fort impact potentiel sur la vie de chacun.

C'est pourquoi il peut être très utile d'acquérir des compétences en matière de statistiques et de lecture de données dès l'école. Non pas tant pour devenir des statisticiens à l'âge adulte, mais pour apprendre à lire et à décoder correctement le monde qui nous entoure.

Parmi les différents projets qui visent à développer la culture statistique dans les écoles, ce travail a le but d' illustrer un parcours didactique interdisciplinaire innovant visant à promouvoir et développer les principes d'une citoyenneté active et consciente dans les écoles, à travers des activités de recherche, l'utilisation de données ouvertes, l'utilisation des technologies de l'information et le suivi civique des financements publics européens et nationaux.

Ce projet, destiné aux collèges et lycées, permet de développer des compétences numériques, statistiques et d'éducation civique, ainsi que d'autres compétences transversales telles que le développement du sens critique et la résolution de problèmes, le travail en équipe, les compétences interpersonnelles et de communication, en les intégrant aux contenus des matières d'étude ordinaires, afin d'aider les élèves à connaître et à communiquer, à l'aide de techniques journalistiques, la manière dont les politiques publiques, et en particulier les politiques de cohésion, interviennent dans les lieux où ils vivent.

L'enseignement, combine des moments d'apprentissage asynchrones typiques des MOOCs (Massive Online Open Courses) avec des activités de facilitation en présence dirigées par les enseignants eux-mêmes (préalablement formés ad hoc), des travaux de groupe et des interactions en ligne avec un'équipe de projet.

Les élèves découvriront donc en classe ce à quoi servent les données avec l'aide de leurs enseignants, dans un voyage informatif intéressant dans le monde des chiffres, qui les aidera à développer leur capacité à comprendre les évolutions sociales, culturelles et environnementales grâce aux données fournies par les statistiques officielles. De plus, ils pourront participer à une

initiative pratique, en s'essayant à la création d'un produit de communication et en mettant à l'épreuve leurs capacités d'analyse et leur créativité.

**Mots-clés.** numératie, culture statistique, surveillance civique, citoyenneté active

## MORE STATISTICAL LITERACY = MORE ACTIVE CITIZENSHIP: DATA AND METHODS

**Abstract**
At the international level, the importance of official statistics is now recognized: they are essential for every citizen to be able to know and critically recognize the context in which they live, and to be able to make decisions, even in situations of uncertainty, based on concrete, shared data.

The national statistical institutes, driven by the conviction that a society in which decisions are taken based on objective data is a society in which democracy reigns, have endeavored to make access to official data free and easily accessible to their citizens. Free access to statistical information is a means of ensuring the progress of a democratic society, but it loses its effectiveness if citizens are unable to read, interpret, and analyze the data. So we were faced with a new challenge: not just disseminating data, but promoting it, getting to know it, and making it understood, so that it is useful to the community, a good that has a strong potential impact on everyone's life.

This is why it can be very useful to acquire statistical and data-reading skills at school. Not so much to become statisticians in adulthood, but to learn to read and decode the world around us correctly.

Among the numerous projects aimed at developing statistical literacy in schools, this work aims to illustrate an innovative interdisciplinary teaching pathway designed to promote and develop the principles of active and conscious citizenship in schools, through research activities, the use of open data, the use of information technologies and the civic monitoring of European and national public funding.

The project, aimed at high schools, develops numerical, statistical, and civic education skills, as well as other cross-curricular skills such as critical thinking and problem-solving, teamwork, interpersonal and communication skills, by integrating them into the content of ordinary study subjects, to help pupils learn about and communicate, using journalistic techniques, how public policies, and in particular cohesion policies, operate in the places where they live.

Teaching is project-based, combining asynchronous learning moments typical of MOOCs (Massive Online Open Courses) with face-to-face facilitation activities led by the teachers themselves (previously trained ad hoc), group work, and online interaction with the project team.

Pupils will therefore discover in class what data is used for, with the help of their teachers, in an interesting and informative journey into the world of figures, which will help them develop their ability to understand social, cultural, and environmental developments thanks to the data provided by official statistics. What's more, they will be able to take part in a practical initiative, trying their hand at creating a communication product and putting their analytical skills and creativity to the test.

**Keywords**. numeracy, statistical literacy, civic mindfulness, active citizenship

# 1. Le développement de la littératie statistique

Selon l'OCDE, dans le cadre de l'évaluation internationale des compétences des adultes, la littératie est définie comme l'aptitude à comprendre et utiliser l'information contenue dans des textes écrits dans différents contextes pour atteindre des objectifs, développer des connaissances et des aptitudes. , tandis que la littératie statistique est définie comme « l'aptitude à comprendre et utiliser les informations statistiques dans la vie courante, à la maison, au travail, et en société, en tant que compétence de base nécessaire à l'exercice de la citoyenneté. »

En mots simples, la littératie des données est la capacité de tirer des renseignements utiles des données. Elle met l'accent sur les compétences nécessaires pour travailler avec les données, y compris les connaissances et les compétences requises pour les lire, les analyser, les interpréter, les visualiser et les communiquer ainsi que pour comprendre comment elles sont utilisées dans la prise de décisions.

La littératie des données signifie également avoir les connaissances et les compétences nécessaires pour bien assurer l'intendance des données, ce qui comprend la capacité d'évaluer la qualité des données, de les protéger et de les sécuriser, et d'en assurer l'utilisation responsable et éthique.

## 1.1 La littératie statistique pour l'exercice de la citoyenneté.

Dans le monde d'aujourd'hui, nous sommes inondés d'informations qui nous parviennent de mille façons (radio, journaux, télévision, internet, etc.), le plus souvent remplies de données statistiques. Des millions des données transitent sur internet, continuellement mises à jour, retravaillées et présentées avec la même apparence de crédibilité. Cela pose des problèmes de connaissance et d'orientation, car le citoyen n'est pas toujours en mesure de discerner et de comprendre ce qui se passe dans l'espace public. L'hétérogénéité des modes de collecte, de traitement et de diffusion des données rend leur lecture encore plus difficile.

Avec l'urgence épidémiologique vécue, la hausse de l'inflation et d'autres questions d'actualité il est devenu évident que la maîtrise des données est essentielle pour comprendre ce qui se passe. La capacité à prendre des décisions dans des situations d'incertitude est indispensable au citoyen conscient qui veut comprendre et contrôler les phénomènes qui l'entourent.

Il est donc nécessaire de développer une alphabétisation visant à l'exercice de compétences critiques et qui ne se limite pas aux matières au sens strict. C'est pourquoi il est nécessaire d'accorder une attention particulière au vocabulaire et à la logique des statistiques. Ce n'est qu'en spécifiant correctement le problème, cette discipline peut aider à trouver des solutions.

. La connaissance est en effet l'arme fondamentale pour construire une société ouverte et inclusive, où personne n'est laissé pour compte. La connaissance est l'élément transversal qui sous-tend le changement

## 1.2 La promotion de la littératie statistique dans les écoles

Il semble donc indispensable d'investir dans l'alphabétisation statistique. La connaissance est en effet l'arme fondamentale pour construire une société ouverte et inclusive, où personne n'est laissé pour compte. La connaissance est l'élément transversal qui sous-tend le le changement. Et quel meilleur champ d'intervention que les écoles ? Les élèves sont en effet à la fois les futurs répondants et les futurs utilisateurs des statistiques.

La promotion de la culture statistique dans les écoles contribue en fait à l'objectif 4, de l'Agenda 2030 : "Faire en sorte que, d'ici à 2030, tous les apprenants acquièrent les connaissances et les compétences nécessaires pour promouvoir le développement durable, notamment par l'éducation au développement et aux modes de vie durables, l'égalité entre les sexes, la promotion d'une culture pacifique et non violente , la citoyenneté mondiale et l'appréciation de la diversité culturelle et de la contribution de la culture au développement durable. " (https://unric.org).

Dès le début des années 1980, en Italie les rédacteurs des programmes ministériels de l'enseignement de base ont introduit des contenus statistiques dans les programmes scolaires afin de favoriser l'acquisition des compétences nécessaires à la compréhension de l'information quantitative. Aujourd'hui, ces compétences sont indispensables Il est donc nécessaire d'instaurer une véritable littératie statistique à l'école.

La statistique peut être utilisée comme un " cheval de bataille " efficace pour rapprocher les élèves des mathématiques et de leur puissante capacité à expliquer et à interpréter le monde de manière critique, en utilisant des données et en s'appuyant sur leurs propres opinions (MIUR 2018).

Mais comment promouvoir la culture et la compréhension statistiques chez les élèves, sans perdre de vue l'objectif de développer l'esprit critique ?

Partant du principe que la réflexion statistique est aussi nécessaire à la citoyenneté active que la capacité à lire et à écrire, ce document relate quelques expériences et initiatives au service de l'enseignement de la statistique comme outil de sensibilisation civique

# 2.Le project: À l'école pour une citoyenneté active”

## 2.1 Structure et méthodes de mise en œuvre

Le project est un parcours didactique interdisciplinaire innovant destiné aux colleges et aux lycées, qui favorise les activités de contrôle civique des financements publics, notamment par l'utilisation de données ouvertes et de technologies de l'information et de la communication.

Le programme d'enseignement du parcours favorise aussi l'acquisition de compétences en matière d'éducation civique, de numérique, de statistiques ,d'éducation civique, de journalisme numérique, et d'autres compétences transversales telles que le développement du sens critique , de la résolution de problèmes, du travail en équipe et des compétences interpersonnelles et de communication, en les intégrant au contenu des matières d'études ordinaires.

L'objectif de chaque équipe d’ètudiants est de mener une recherche thématique pour étudier les caractéristiques socio-économiques, environnementales et/ou culturelles de son territoire à partir d'une intervention financée par les politiques de cohésion sur un thème d'intérêt, choisi sur la base d'informations publiées en format ouvert sur le portail européen vérifiant ainsi comment les politiques publiques interviennent pour améliorer le contexte local.

Le projet est divisé en 4 leçons :

● Planification (leçon 1). Apprendre en quoi consiste le suivi civique, choisir sur le site d' portail de l'Union européenne un projet financé dans votre région à surveiller, identifier une question de recherche. Identifiez une question de recherche, formez le groupe de travail en classe et répartissez-vous en fonction des rôles. Recherchez plus d'informations sur le projet Si le projet est choisi,

reconstituez le processus administratif et les décisions publiques qui ont déterminé le projet, identifier les acteurs publics et privés impliqués dans sa mise en œuvre.

● Analyser (leçon 2). Apprendre les techniques de recherche quantitative et qualitative, comprendre ce qu'est l'open data et rechercher des données en rapport avec le sujet choisi, construire un indicateur avec les données trouvées, comprendre le flux de travail. du journalisme de données.

● Explorer (leçon 3). Explorer sur le terrain les progrès du projet choisi par le biais d'une visite. d'un suivi sur place, d'entretiens avec les responsables de la mise en œuvre, de rencontres avec les institutions. Rédiger un rapport de suivi détaillé

● Raconter des histoires (leçon 4). Apprentissage des techniques de communication, conception et mise en œuvre d'une campagne. de sensibilisation et d'engagement pour illustrer les résultats du suivi civique. Organiser un événement public et impliquer la communauté cible pour poursuivre la surveillance sur le projet choisi.

Les classes participant au projet sont soutenues par l'enseignant de référence, avec la collaboration éventuelle d'un enseignant de soutien. Les enseignants et les élèves participent au projet avec le soutien des réseaux territoriaux du project, composés des centres Europe Direct et CDE, des organisations "Amis", c'est-à dire des associations à but non lucratif actives dans le domaine des politiques de cohésion, des personnes de contact territoriales de l'Institut national de la statistique, et tous concourent pour des prix et des récompenses, y compris des voyages éducatifs à Bruxelles auprès des institutions européennes ainsi que des opportunités de formation et d'expérience, en ligne et en présence, et bien plus encore grâce au soutien des nombreux partenaires du projet. Les régions et les collectivités partenaires apportent leur soutien dans la liaison avec les institutions nationales et locales, avec les autres partenaires du projet (entreprises, fondations, associations) et dans les activités de valorisation du projet à l'échelle nationale**.**

# Bibliographie


OECD (2018) Numeracy practices and numeracy skills among adults https:\\www.oecd.org\officialdocuments\publicdisplaydocumentpdf\?cote=EDU\WKP(2018)13&doclanguage=en 5

Carpentieri, J., J. Lister et L. Frumkin (2009), « Adult numeracy: a review of research », http://roar.uel.ac.uk/1340

Hanushek, E. et al. (2015), « Returns to skills around the world: Evidence from PIAAC », European Economic Review, vol. 73, pp. 103-130, http://dx.doi.org/10.1016/j.euroecorev.2014.10.006

Desjardins, R. et K. Rubenson (2011), « An Analysis of Skill Mismatch Using Direct Measures of Skills », Documents de travail de l'OCDE sur l'éducation, no. 63, Éditions OCDE, Paris, http://dx.doi.org/10.1787/5kg3nh9h52g5-en

Istat https:\\olimpiadistatistica.istat.it\pluginfile.php\418696\mod_resource\content\5\catalogo.pdf Dati alla mano https://www.istat.it/it/dati-alla-mano

Coesione Italia https://opencoesione.gov.it/it/ASOC/

Centre régional d'information des Nations Unies https://unric.org

OCDE Skills Surveys https://www.oecd.org/skills/piaac/piaacdesign

UNICEF (2016), Mapping the Global Goals for Sustainable Development and the Convention on the Rights of the Child. https://www.unicef.org/documents/mapping-global-goals-sustainable-development-andconvention-rights-child.

Von Glasersfeld E. (1987), The Construction of Knowledge, Contributions to Conceptual Semantics. Salinas:Intersystems Publications.

Wallmann K. (1993), Enhancing Statistical Literacy: Enriching Our Society, Journal of the American Statistical Association, 88, 421: 1-8.

# Guy Brousseau et la statistique

Marthe-Aline Jutand*†[1] and Bernard Sarrazy*‡

[1]Cultures et Diffusion des Savoirs – Université de Bordeaux – France

**Résumé**

La communauté de statisticiens et statisticiennes souhaite rendre hommage à Guy Brousseau qui est décédé à l'âge de 91 ans en février 2024. Connu comme un des fondateur de l'école française de recherche en didactique des mathématiques, et premier récipiendaire de la médaille Felix Klein, crée par l'ICMI en 2003, ses recherches et expérimentations dans le champ de l'enseignement de la probabilité et de la statistique sont moins diffusées. Et pourtant il a produit un certain nombre de travaux en didactique de la statistique et d'expériences au sein de l'école primaire Michelet de Talence, qui sont de réelles pistes d'inspiration pour l'enseignement de la statistique du primaire au supérieur.
Lors des 41◦JdS de la SFdS qui se sont déroulées à Bordeaux en 2009, Guy Brousseau disait " *Il y a plusieurs millénaires que l'on essaie de prévoir l'avenir (et que l'on exploite l'impossibilité de le faire) en interprétant le passé et ses statistiques. Notre culture est chargée des traces de ces efforts et des résidus de leurs échecs. Il y a très peu de temps que ces deux champs sont devenus des sciences et des objets de théories mathématiques. Le travail spontané de transposition didactique est lent, chaotique et il procure un accès assez sélectif aux connaissances. Les connaissances de statistique doivent être vite et mieux diffusées. Les solutions à ce problème didactique sont actuellement recherchées dans des sciences qui apportent des renseignements intéressants mais très indirects comme la psychologie ou les neurosciences et qui ne prennent pas le processus spécifique de la création ou de la recréation d'une connaissance précise comme objet d'études théoriques et expérimentales. On peut espérer que la Didactique, en évitant les inférences douteuses auxquelles ces approches latérales nous condamnent, contribuera à améliorer la diffusion de la statistique.*"

C'est donc pour continuer à contribuer à améliorer la diffusion et l'éducation à la statistique qu'il est important de ne pas oublier les travaux de Guy Brousseau et de présenter ses propositions pédagogiques pour initier les enfants à la démarche probabiliste et statistique. Pour appréhender cela il est important de reposer son cadre de pensée (théorie des situations ...) et de décrire des situations pédagogiques qu'il a élaborées et qui peuvent donner des pistes de réflexions aux enseignants.

Référence - Guy Brousseau. Alternatives en didactique de la statistique. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France

**Mots-Clés:** didactique de la statistique

*Intervenant

†Auteur correspondant: marthe-aline.jutand@u-bordeaux.fr

‡Auteur correspondant: bernard.sarrazy@u-bordeaux.fr

# Session "Trucs et Astuces" pour Statistique Mathématique

# About the van Trees inequality and its use for statistical lower bounds.

Élisabeth Gassiat*[1] and Gilles Stoltz

[1]Laboratoire de Mathématiques d'Orsay – Université Paris-Saclay, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR8628 – France

**Résumé**

In this talk, I will present the van Trees inequality and how it can be used to derive lower bounds on the minimax quadratic risk for parametric, semi parametric and non parametric problems. In particular, I will provide an elementary proof of the local asymptotic minimax theorem for quadratic loss functions, avoiding the beautiful but sophisticated Hajek-Le Cam theory of convergence of experiments.

**Mots-Clés:** inégalité d'information, borne inférieure

*Intervenant

# Tests d'indépendance et tests d'homogénéité basés sur des méthodes à noyaux

Mélisande Albert [1], Béatrice Laurent [1], Amandine Marrel [2], Anouar Meynaoui [3], Antonin Schrab [4], Ilmun Kim [5], Benjamin Guedj [4] & Arthur Gretton [6]

[1] *Institut de Mathématiques de Toulouse, INSA Toulouse, melisande.albert@insa-toulouse.fr, beatrice.laurent@insa-toulouse.fr*
[2] *CEA, DES, IRESNE, DER, SESI, Cadarache, associée à l'IMT, amandine.marrel@cea.fr*
[3] *Institut de Recherche Mathématique de Rennes, Université de Rennes 2, anouar.meynaoui@univ-rennes2.fr*
[4] *Centre for Artificial Intelligence, University College London & Inria London, a.schrab@ucl.ac.uk, b.guedj@ucl.ac.uk*
[5] *Department of Statistics & Data Science, Yonsei University, ilmun@yonsei.ac.kr*
[6] *Gatsby Computational Neuroscience Unit, University College London, arthur.gretton@gmail.com*

**Résumé.** Cet exposé s'appuie sur deux articles respectivement en collaboration avec M. Albert, A. Marrel et A. Meynaoui [Albert et al., 2022] et avec A. Schrab, I. Kim, M. Albert, B. Guedj et A. Gretton [Schrab et al., 2023]. Nous nous intéressons d'une part à tester l'indépendance de deux vecteurs $X \in \mathbb{R}^p$ et $Y \in \mathbb{R}^q$ à partir de l'observation d'un $n$-échantillon $((X_1, Y_1), \dots, (X_n, Y_n))$ et d'autre part à tester que deux échantillons indépendants de variables aléatoires à valeurs dans $\mathbb{R}^p$, $(X_1, \dots, X_m)$ i.i.d. de loi de probabilité $P$ et $(Y_1, \dots, Y_n)$ i.i.d. de loi de probabilité $Q$, ont même loi. Le point commun de ces deux papiers est d'utiliser la notion de MMD (Maximum Mean Discrepancy) qui définit une métrique entre lois de probabilités basée sur des noyaux dans des espaces de Hilbert à noyau reproduisant (RKHS). Plus précisément, étant donné un RKHS $\mathcal{H}_k$ associé au noyau $k$, la MMD entre deux mesures de probabilités $P$ et $Q$ est définie par

$$\mathrm{MMD}(P, Q, \mathcal{H}_k) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim P}\left[f(X)\right] - \mathbb{E}_{Y \sim Q}\left[f(Y)\right].$$

Pour certains types de noyaux, dits caractéristiques, la nullité de $\mathrm{MMD}(P, Q, \mathcal{H}_k)$ équivaut à l'égalité des mesures de probabilités $P$ et $Q$.

Pour le problème de test d'égalité des lois $P$ et $Q$ de deux échantillons, nous nous concentrons sur l'estimation de la quantité $\mathrm{MMD}(P, Q, \mathcal{H}_k)$, pour un certain choix de noyau $k$, en suivant les travaux précurseurs de [Gretton et al., 2007]. Par ailleurs, pour tester l'indépendance de deux vecteurs aléatoires $X$ et $Y$, nous proposons d'utiliser le critère d'indépendance de Hilbert-Schmidt (HSIC) qui a été introduit par [Gretton et al., 2005]), et qui n'est autre que la MMD (associée à un certain noyau) entre la loi du couple $(X, Y)$ et le produit des lois marginales.

L'objectif de l'exposé sera de montrer comment on peut construire des estimateurs de la MMD et du HSIC puis d'en déduire des tests d'homogénéité et des tests d'indépendance.

Nous verrons en particulier le recours à des techniques de permutation pour garantir le niveau des tests. Par ailleurs, nous donnerons des résultats de puissance pour les tests, qui s'appuient sur des inégalités exponentielles pour les U-statistiques dues à [Arcones and Giné, 1993] et [Giné et al., 2000]. Nous discuterons également du choix des noyaux utilisés et nous montrerons l'intérêt d'agréger des tests associés à différents noyaux.

**Mots-clés.** Tests non paramétriques, Maximum Mean Discrepancy, critère d'indépendance de Hilbert-Schmidt, méthodes de permutation, tests agrégés.

**Abstract.** This presentation is based on two papers respectively in collaboration with M. Albert, A. Marrel and A. Meynaoui [Albert et al., 2022] as well as A. Schrab, I. Kim, M. Albert, B. Guedj and A. Gretton [Schrab et al., 2023]. We are interested, on the one hand, in testing the independence of two vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ from the observation of an $n$-sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ and, on the other hand, in testing that two independent samples of random variables with values in $\mathbb{R}^p$, $(X_1, \ldots, X_m)$ i. i.d. with probability distribution $P$ and $(Y_1, \ldots, Y_n)$ i.i.d. with probability distribution $Q$, have the same distribution. These two papers have in common their use of the notion of MMD (Maximum Mean Discrepancy), which defines a metric between probability distributions based on kernels in reproducing kernel Hilbert spaces (RKHS). More precisely, given an RKHS $\mathcal{H}_k$ associated with kernel $k$, the MMD between two probability measures $P$ and $Q$ is defined by

$$\mathrm{MMD}(P, Q, \mathcal{H}_k) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim P}\left[f(X)\right] - \mathbb{E}_{Y \sim Q}\left[f(Y)\right].$$

For certain types of kernels, known as characteristic kernels, the nullity of $\mathrm{MMD}(P, Q, \mathcal{H}_k)$ is equivalent to the equality of the probability measures $P$ and $Q$.

For the problem of testing the equality of the laws $P$ and $Q$ of two samples, we focus on estimating the quantity $\mathrm{MMD}(P, Q, \mathcal{H}_k)$, for a certain choice of kernel $k$, following the seminal work of [Gretton et al., 2007]. Furthermore, to test the independence of two random vectors $X$ and $Y$, we propose to use the Hilbert-Schmidt independence criterion (HSIC) introduced by [Gretton et al., 2005]), which is none other than the MMD (associated with a certain kernel) between the law of the pair $(X, Y)$ and the product of the marginal laws.

The aim of this presentation will be to propose estimators of the MMD and HSIC, and then to derive tests of homogeneity and independence. In particular, we will explain how the use of permutation techniques allows to guarantee the level of the tests. In addition, we will give power results for the tests, which are based on exponential inequalities for U-statistics due to [Arcones and Giné, 1993] and [Giné et al., 2000]. We will also discuss the choice of kernels used and show the benefits of aggregating tests associated with different kernels.

**Keywords.** Nonparametric tests, Maximum Mean Discrepancy, Hilbert-Schmidt Independence Criterion, permutation methods, aggregated tests.

# Références

[Albert et al., 2022] Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on HSIC measures. *Ann. Statist.*, 50(2) :858–879.

[Arcones and Giné, 1993] Arcones, M. A. and Giné, E. (1993). Limit theorems for $u$-processes. *Ann. Probab.*, 21(3) :1494–1542.

[Giné et al., 2000] Giné, E., Latał a, R., and Zinn, J. (2000). Exponential and moment inequalities for $U$-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA.

[Gretton et al., 2007] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2007). A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems*, pages 513–520, Cambridge, MA. MIT Press.

[Gretton et al., 2005] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In Jain, S., Simon, H. U., and Tomita, E., editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Schrab et al., 2023] Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD aggregated two-sample test. *J. Mach. Learn. Res.*, 24 :Paper No. [194], 81.

# Statistique appliquée à la médecine 2

# Test allocation based on risk of infection from first and second order contact tracing

Gabriela Bayolo Soler [1] & Miraine Dávila Felipe [2] & Ghislaine Gayraud [3]

[1] *Université de technologie de Compiègne - France - gabriela.bayolo-soler@utc.fr*
[2] *Université de technologie de Compiègne - France - miraine.davila-felipe@utc.fr*
[3] *Université de technologie de Compiègne - France - ghislaine.gayraud@utc.fr*

**Résumé.** Face à une quantité limitée des ressources disponibles, les stratégies visant à atténuer la propagation d'une épidémie, telles que les tests aléatoires et la recherche des contacts, deviennent inefficaces. Nous proposons ici d'allouer les ressources de manière plus efficace, en calculant le risque individuel d'infection au cours du temps, basé sur l'observation partielle de la propagation de l'épidémie sur un réseau de contacts ; ce risque est défini comme la probabilité d'être infecté par n'importe quelle chaîne de transmission possible jusqu'à une longueur de deux, provenant d'individus récemment détectés. Pour évaluer les performances de notre méthode et les effets de certains paramètres clés, nous effectuons des expériences comparatives simulées en utilisant des données générées par un modèle basé sur des agents.

**Mots-clés.** réseaux de contacts, recherche des contacts, atténuation des épidémies, probabilité d'infection, estimation des risques, attribution des tests

**Abstract.** Under limited available resources, strategies for mitigating the propagation of an epidemic such as random testing and contact tracing become inefficient. Here, we propose to accurately allocate the resources by computing over time an individual risk of infection based on the partial observation of the epidemic spreading on a contact network; this risk is defined as the probability of getting infected from any possible transmission chain up to length two, originating from recently detected individuals. To evaluate the performance of our method and the effects of some key parameters, we carry out comparative simulated experiments using data generated by an agent-based model.

**Keywords.** contact networks, contact tracing, epidemic mitigation, probability of infection, risk estimation, test allocation

# Introduction

In this work, we propose a method to compute the risk of infection of individuals in the population over time, based on the partial observation of the epidemic spreading through the population contact network. The risk of each individual is defined as her/his (marginal) probability of infection conditionally on the observed variables in the recent past, and the higher-risk individuals can get notified to be tested, quarantined, or applied any other preventive measures. Thus, the quantification of the infection risk is proposed here as a tool

to allocate the available resources more rationally than just randomly. Similar intervention approaches have been shown to have a positive impact on the mitigation of epidemics, see for instance Baker et al. (2021); Herbrich et al. (2022); Batlle et al. (2022); Romijnders et al. (2023); Guttal et al. (2020); Bestvina and Thornton (2021); Murphy et al. (2021); Gupta et al. (2023); Bengio et al. (2021); Sattler et al. (2020); Alsdurf et al. (2020).

Among this recent research with the same aim as ours, in Bestvina and Thornton (2021) and Batlle et al. (2022), the risk is computed using Monte Carlo methods. In Bestvina and Thornton (2021), the authors estimate the individual infection probability up to 3° contact tracing, arguing that it improves the detection of asymptomatic patients in diseases with a high percentage of them. Here, instead, we derive an explicit formula for these probabilities taking into account up to 2° contacts, and hence avoiding the large computing power and the centralised information required by Monte Carlo methods. Other works such as Baker et al. (2021) and Guttal et al. (2020) avoid the use of Monte Carlo methods by using the mean-field approximation to evaluate the individual risk of infection. However, the way the risk is propagated is "bidirectional", meaning that it is not only "forward" in the direction of the transmission given the observations. Indeed, they suppose that individuals interchange their risk information at each time step if there is an edge between them, regardless of the previous path followed by the transmitted risk. Despite the similarities in the use of the MF hypothesis, it should be noticed that in Guttal et al. (2020) the network and propagation model are simpler than in Baker et al. (2021). The latter deals with more realistic models, including the OpenABM-Covid19 model used in our simulations. Moreover, in Baker et al. (2021) a second method is developed, that estimates the individual infection risk as the posterior distribution conditional on the test observations through the Belief Propagation inference algorithms. Similar computations are achieved in Herbrich et al. (2022) and Romijnders et al. (2023) using Gibbs Sampling and Factorized Neighbors respectively. Another related methodology is presented in a series of works Bengio et al. (2021); Gupta et al. (2023); Alsdurf et al. (2020); Biazzo et al. (2022); Shah et al. (2020); Čutura et al. (2021); Tomy et al. (2022); Tan et al. (2023), where the authors achieve the risk computation using deep learning algorithms based on neural networks.

We propose and simulate the following mitigation strategy: every day the probability of being infected is computed for the individuals at risk, and a fixed number ($\eta$, number of daily available tests) of the highest-ranked individuals are tested; the newly detected individuals are put in quarantine the day after, and the process is repeated each day during an intervention period.

We briefly describe our approach in Section 1 and we present some results in Section 2.

# 1 Methods

To be more precise about our approach, we suppose that the infectious statuses of individuals are (partially) observed through testing, as well as the underlying contact network. Actually, we consider *at risk* not only the first-degree contacts of detected individuals (1° contacts) but also their subsequent contacts (2° contacts). In the sequel, to distinguish the different

groups of individuals under consideration, we call *index cases* the detected individuals who are infectious, 1° *contacts* the individuals who are not detected but interacted with index cases while the latter were infectious, and, 2° *contacts* the individuals who are not detected but interacted with 1° contacts after the latter were in contact with index cases. In addition, 1° *interaction* and 2° *interaction* refer to the risky encounter between index cases and 1° contacts, and between 1° and 2° contacts respectively. Then, we compute the probability for each individual at risk of having been infected by one of the index cases in the previous days (fixed time window), through chains of transmission of length one or two.

*Infectious disease spreading on the network*

We consider a population consisting of $N$ ($N \in \mathbb{N}$) individuals that stays constant over time. At any discrete time $t$ ($t \in \mathbb{N}$), the social structure (interactions between individuals at $t$) is represented by an undirected graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$, corresponding to the set of vertices $\mathcal{V}$ (the individuals) and the set of the edges $\mathcal{E}_t$ (the interactions at time $t$ between the corresponding individuals). Here, we consider an individual-based SIR dynamic spreading on the underlying social network. The possible individual statuses are only Susceptible ($S$), Infected ($I$) and Removed ($R$), and the only possible status evolution over time are $S \to I$ and $I \to R$, where $R$ considered as an absorbing state. We denote by $X_t^i \in \{S, I, R\}$ the random variable corresponding to the status of individual $i$ at time $t$.

*Observations*

At a given time $t \geq 0$, the set of observations $\mathcal{O}_t$ is provided by the graph of interactions during the recent days and the set of individuals with a positive and a negative result.

## Risk of infection via transmission chains

For any time $t$ and any individual $j$ at risk, our aim is to estimate the probability of infection of $j$ given the set of observations $\mathcal{O}_t$, that is

$$\mathbb{P}\left(X_t^j = I \mid \mathcal{O}_t\right) \tag{1}$$

In order to compute this probability we propose two approaches based on two different degrees of interactions. To differentiate both methods, we call them in the sequel 1°contact tracing (1°CT) and 2°contact tracing (2°CT). In the first approach, the risk of infection is based on 1° interactions, while the second proposes a more accurate risk of infection, defined from both 1° and 2° interactions.

We illustrate, using the toy example in Figure 1, the fundamental differences between 1°CT and 2°CT method. The Figure shows the interactions between 3 individuals during 4 days. At day 4, individual $a$ is tested positive and then their contacts in the past are traced and ranked. For 1°CT is considered only individual $b$ at risk, while 2°CT considers both individuals $b$ and $c$.

We introduce a truncation parameter $\gamma \in \mathbb{N}$ corresponding to the infection time-frame of interest. More precisely, for any individual not tested positive, we are interested in the


(a) $t = 1$ (b) $t = 2$ (c) $t = 3$ (d) $t = 4$

Figure 1: Temporal interaction networks in a population of $N = 3$. Individual $a$ is detected at day $t = 4$.

approximate probability of being infected in the interval $[t-\gamma : t]$ given the set of observations at time $t$. For 2°CT method, we also introduce the parameter $\zeta$ ($\zeta \geq \gamma$) as the time-frame for the infection date of the 1°contacts, meaning that it lies in the interval of time $[t - \zeta : t - 1]$. As a consequence, we are interested in 1° interactions that occur in $[t - \zeta : t - 1]$ and in the 2° interactions that occur after a possible transmission due to a 1° interaction in the interval of time $[t - \gamma : t]$.

# 2 Simulation results

The simulation starts at time $t = 0$. At the beginning, all individuals are susceptible ($S$), except for a small number $N_0$ of infectious individuals ("patients zero"). Every day, starting from $t = 1$, a proportion $p_s$, respectively $p_m$, of individuals with newly developed severe and mild symptoms are tested, detected and quarantined. Later, at a fixed date $t_0$ in $[1 : T]$, the intervention starts, and it is carried out daily until the mitigation of the epidemic or the end of the study. At any $t \geq t_0$, the intervention strategy based on the 2°CT method consists of tracing 1° and 2°contacts, computing their risk of infection, and ranking them according to their risk values. Then, the first $\eta$ individuals in the ranking are tested, and the newly detected ones become index cases and are quarantined. The default quarantine protocol stops the interactions in the occupation and random network, but those within the household are maintained. The tests are assumed to be perfect, and we suppose that the test results are available the same day on which the tests are performed. The intervention related to the 1°CT method is analogous to the one for the 2°CT method, except that only the 1°contacts are traced and ranked.

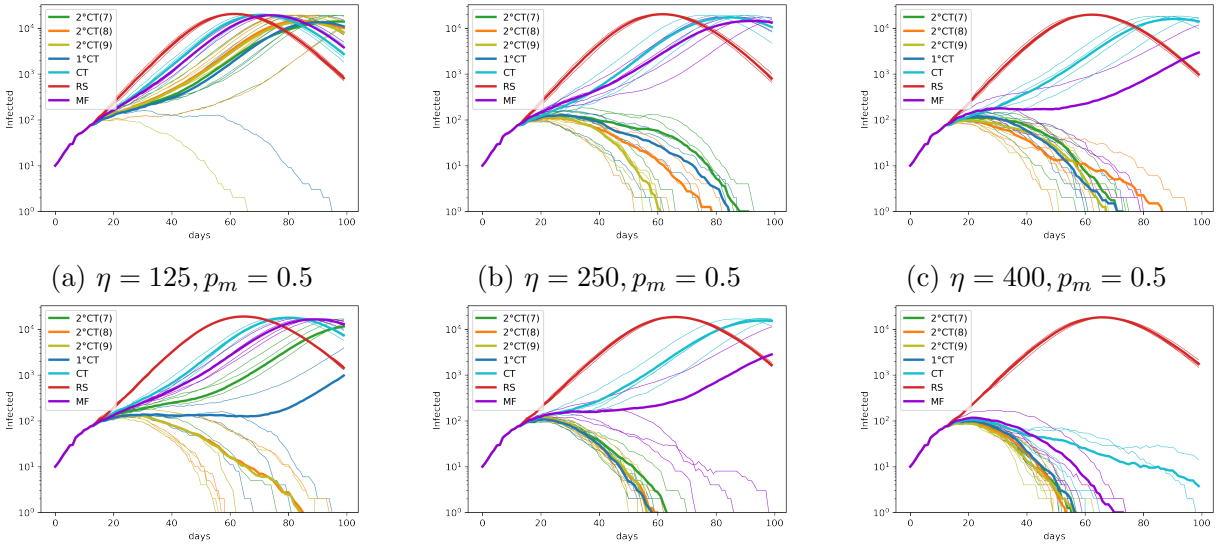
To test our proposed method on a proper data set, we generate the data using the OpenABM-Covid19 model introduced by Hinch et al. (2021).

## Results

In this section, we present the results obtained using the intervention based on the risk, through the simulation of different scenarios. For all the simulations, the propagation of the epidemic is identical until $t_0$, while it might change after $t_0$, when the intervention method starts, depending on the particular scenario. In the figures each thin line represents the result obtained for the realisation associated with one *seed*, while the thick lines correspond to the average of all the realisations.

To evaluate the efficiency of the proposed 1°CT and 2°CT methods to mitigate an epi-

demic, we compare them with three other ranking strategies: Random Selecting (RS, individuals are ranked randomly), Contact Tracing (CT, individuals are ranked according to their number of interactions with detected individuals in the time-frame $[t-\gamma : t]$) and Mean-Field (MF, individuals are ranked according to the mean-field risk approximation presented in Baker et al. (2021)).




(a) $\eta = 125, p_m = 0.5$ (b) $\eta = 250, p_m = 0.5$ (c) $\eta = 400, p_m = 0.5$

(d) $\eta = 125, p_m = 0.75$ (e) $\eta = 250, p_m = 0.75$ (f) $\eta = 400, p_m = 0.75$

Figure 2: Effect of the parameters $\eta$ (the number of daily available tests, increasing from left to right) and $p_m$ (the proportion of daily detected individuals with mild symptoms, increasing from top to bottom) on the epidemic spreading for the strategies 1°CT, 2°CT, CT, RS and MF. In all simulations we consider $T = 100$, $N = 50K$, $t_0 = 12$, $N_0 = 10$ and $p_S = 1$. We fix the parameters $\gamma = 6$ and $\zeta = 7, 8, 9$ (indicated in the legend as 2°CT(7), 2°CT(8), 2°CT(9), respectively). The values for the parameters in MF strategy are $\rho_{MF} = 5$ and $t_{MF} = 10$.

We compare the five strategies in Figure 2, in which we display the number of infectious individuals in logarithmic scale through time across a broad range of values for the parameters. In particular, we increase the number of daily available tests from the left panels to the right ones, and we increase the proportion of daily detected mild symptomatic individuals from top to bottom. As expected for all strategies, a higher value of $p_m$ and/or $\eta$ improves the mitigation of the epidemic in terms of the duration and the total number of infected individuals. The simulations show that our proposed methods (1°CT and 2°CT) improve considerably the results compared to the MF and the usual CT, which are all better than the RS strategy. The latter does not mitigate the epidemic even with a high number of daily available tests and a high value of $p_m$, while the MF and CT methods achieve the mitigation for a large value of $\eta$. We also study the 2°CT method for different time-frames in which the 1°contact can get infected, that is in $[t-\zeta : t]$, where we consider $\zeta = 7$ in green, $\zeta = 8$ in orange and $\zeta = 9$ in yellow. Figure 2 shows that the results are improved as $\zeta$ increases. In particular, it should be noticed that the 2°CT method with $\zeta = 8$ and $\zeta = 9$ gives better results than the 1°CT method. However, the 1°CT method requires less individual information and therefore it is better in terms of privacy restrictions. From these

results, a trade-off can arise between getting better results with computationally demanding (2°CT method) and preserving individual privacy with simpler and faster computation. Indeed, it is worth mentioning that for a high enough number of daily available tests and/or a high enough proportion of mild observed, the methods 1°CT and 2°CT have similar effects on the mitigation of the epidemic; hence in this case, we recommend the use of the 1°CT method than the 2°CT method.

# Conclusions

A crucial aspect of our work is that we consider a rather realistic contact network and a detailed disease spread model, see Hinch et al. (2021). Another significant aspect is that we consider 2° contact tracing, providing a more accurate estimation of the risk compared with the 1° contact tracing. Although our approach can be extended to 3° contacts and beyond, the calculations would get much heavier, and we argue that the gain in the effectiveness of the mitigation would not be significant, due to the uncertainty on the statuses of the intermediate individuals in the chains of transmission. One more core feature of our method is that to compute the risk, neither the whole contact network nor a centralised setup (contacts and individual information) is required. These characteristics are essential for the practical implementation using digital contact tracing applications, where the interchange of information between contacts over the whole network could be difficult due to a huge amount of personal data impacted by privacy restrictions. These difficulties can be intensified in the centralised case, see Romijnders et al. (2023). Furthermore, compared to the previously inference algorithms in the literature used to calculate the individual risk of being infected (Belief Propagation, Gibbs Sampling and Factorized Neighbors), our risk calculation is simpler: while these algorithms integrate the observations at any time $t$ by updating and re-propagating the risks step-by-step in a given time interval previous to $t$ for every contact (up to any contact degree) of all the individuals in the population, we calculate directly the risk at $t$ of individuals in contact (up to 2°) with someone detected by integrating the probability of any possible path of length up to 2 that might lead to the infection of these individuals. In this way, we avoid any cycling back phenomenon, and we do not need to update the risk of all individuals for every time step in the contact tracing time window, getting a very low level of messages interchange between individuals, see Romijnders et al. (2023).

# Further information

The repository with the code is available in https://github.com/gbayolo26/risk_estimation. The presentation will be based on the recent submitted paper Bayolo Soler et al. (2023).

# References


Alsdurf, H., Belliveau, E., Bengio, Y., Deleu, T., Gupta, P., Ippolito, D., Janda, R., Jarvie, M., Kolody, T., Krastev, S., Maharaj, T., Obryk, R., Pilat, D., Pisano, V., Prud'homme, B., Qu, M., Rahaman, N., Rish, I., Rousseau, J.-F., Sharma, A., Struck, B., Tang, J., Weiss, M. and Yu, Y. W. (2020) Covi white paper.

Baker, A., Biazzo, I., Braunstein, A., Catania, G., Dall'Asta, L., Ingrosso, A., Krzakala, F., Mazza, F., Mé zard, M., Muntoni, A. P., Refinetti, M., Mannelli, S. S. and Zdeborová, L. (2021) Epidemic mitigation by statistical inference from contact tracing data. *Proceedings of the National Academy of Sciences*, **118**, e2106548118.

Batlle, P., Bruna, J., Fernandez-Granda, C. and Preciado, V. M. (2022) Adaptive test allocation for outbreak detection and tracking in social contact networks. *SIAM Journal on Control and Optimization*, **60**, S274–S293.

Bayolo Soler, G., Dávila Felipe, M. and Gayraud, G. (2023) Test allocation based on risk of infection from first and second order contact tracing. *arXiv preprint arXiv:2311.02094*.

Bengio, Y., Gupta, P., Maharaj, T., Rahaman, N., Weiss, M., Deleu, T., Muller, E. B., Qu, M., Schmidt, V., St-Charles, P.-L., Alsdurf, H., Bilaniuk, O., Buckeridge, D., Marceau-Caron, G., Carrier, P.-L., Ghosn, J., Ortiz Gagne, S., Pal, C., Rish, I., Schölkopf, B., Sharma, A., Tang, J. and Williams, A. (2021) Predicting infectiousness for proactive contact tracing. In *9th International Conference on Learning Representations (ICLR)*. Virtual Conference.

Bestvina, I. and Thornton, W. (2021) Contact tracing infection risk estimate. Distributed simulation aproach. URL: `https://www.viratrace.org/#/`.

Biazzo, I., Braunstein, A., Dall'Asta, L. and Mazza, F. (2022) A bayesian generative neural network framework for epidemic inference problems. *Scientific Reports*, **12**, 19673.

Gupta, P., Maharaj, T., Weiss, M., Rahaman, N., Alsdurf, H., Minoyan, N., Harnois-Leblanc, S., Merckx, J., Williams, A., Schmidt, V. et al. (2023) Proactive contact tracing. *PLOS Digital Health*, **2**, e0000199.

Guttal, V., Krishna, S. and Siddharthan, R. (2020) Risk assessment via layered mobile contact tracing for epidemiological intervention. *medRxiv*. URL: `https://www.medrxiv.org/content/early/2020/05/01/2020.04.26.20080648`.

Herbrich, R., Rastogi, R. and Vollgraf, R. (2022) CRISP: A Probabilistic Model for Individual-Level COVID-19 Infection Risk Estimation Based on Contact Data.

Hinch, R., Probert, W. J., Nurtay, A., Kendall, M., Wymant, C., Hall, M., Lythgoe, K., Bulas Cruz, A., Zhao, L., Stewart, A. et al. (2021) OpenABM-Covid19—an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLoS computational biology*, **17**, e1009146.

Murphy, K., Kumar, A. and Serghiou, S. (2021) Risk score learning for COVID-19 contact tracing apps. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, vol. 149 of *Proceedings of Machine Learning Research*, pp. 373–390. PMLR. URL: https://proceedings.mlr.press/v149/murphy21a.html.

Romijnders, R., Asano, Y. M., Louizos, C. and Welling, M. (2023) No time to waste: practical statistical contact tracing with few low-bit messages. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, vol. 206 of *Proceedings of Machine Learning Research*, 7943–7960. PMLR. URL: https://proceedings.mlr.press/v206/romijnders23a.html.

Sattler, F., Ma, J., Wagner, P., Neumann, D., Wenzel, M., Schäfer, R., Samek, W., Müller, K.-R. and Wiegand, T. (2020) Risk estimation of SARS-CoV-2 transmission from bluetooth low energy measurements. *NPJ Digital Medicine*, **3**, 129. URL: https://doi.org/10.1038/s41746-020-00340-0.

Shah, C., Dehmamy, N., Perra, N., Chinazzi, M., Barabási, A.-L., Vespignani, A. and Yu, R. (2020) Finding patient zero: Learning contagion source with graph neural networks.

Tan, C. W., Yu, P.-D., Chen, S. and Poor, H. V. (2023) Deeptrace: Learning to optimize contact tracing in epidemic networks with graph neural networks.

Tomy, A., Razzanelli, M., Di Lauro, F., Rus, D. and Della Santina, C. (2022) Estimating the state of epidemics spreading with graph neural networks. *Nonlinear Dynamics*, **109**, 249–263.

Čutura, G., Li, B., Swami, A. and Segarra, S. (2021) Deep demixing: Reconstructing the evolution of epidemics using graph neural networks. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 2204–2208.

# Implementation d'un modéle de progression multivarié (Leaspy) pour l' étude de l'évolution et l'identification de sous-groupes sur la maladie de CADASIL

Sofia Kaisaridi [1], Hugues Chabriat [2] & Sophie Tezenas du Montcel [3]

[1] *Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France, sofia.kaisaridi@icm-institute.org*

[2] *Centre Neurovasculaire Translationnel-Centre de Référence CERVCO, FHU NeuroVasc, Hôpital Lariboisière, AP-HP, Université de Paris, INSERM, Unité Mixte de Recherche 1161, Paris, France, hugues.chabriat@aphp.fr*

[3] *Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France, sophie.tezenas@aphp.fr*

**Résumé.** Les modèles de progression des maladies sont un outil prometteur pour analyser des données longitudinales présentant de multiples modalités. De tels modèles peuvent être utilisés pour estimer une progression de la maladie à long terme et reconstruire des trajectoires individuelles qui peuvent rendre compte de la variabilité entre les patients, mais aussi entre les modalités. Nous avons implémenté la version multivariée, combinant 14 tests cliniques, d'un modèle bayésien à effets mixtes (Leaspy), qui implique une reparamétrisation temporelle pour analyser l'évolution d'une artériopathie cérébrale autosomique dominante à l'origine d'infarctus sous-corticaux et d'une leucoencéphalopathie (CADASIL). Nous avons analysé les données de 395 patients avec 2007 visites, recrutés au Centre National de Référence français CERVCO, en utilisant Leaspy pour évaluer la variabilité temporelle, résultant de différents rythmes de progression et de décalages temporels, ainsi que la variabilité spatiale, prenant la forme d'une modification dans la séquence des événements. Notre analyse multivariée nous a permis de modéliser la progression de la maladie comme un vecteur à 14 dimensions, et ainsi de rendre compte des relations entre les variables dépendantes, qui ont montré des variations dans le point temporel de début de détérioration. Nous avons ensuite implémenté un modèle de mélange gaussien (GMM) pour les paramètres individuels afin d'identifier des sous-types sous-jacents de la maladie. Nous avons identifié un sous-groupe de patients avec un début précoce et une progression rapide, présentant des symptômes moteurs et des déficits neurologiques plus précoces, et un groupe avec un début tardif et une progression lente, montrant des symptômes cognitifs plus précoces. Pour conclure le modèle multivarié parvient à mettre en évidence les différences dans l'évolution des scores cliniques, suggérant un déclin progressif mais également hétérogène dans CADASIL et le profilage a révélé deux sous-types impliquant à la fois la variabilité temporelle et spatiale.

**Mots-clés.** Modèle de progression des maladies, Multivariée, Bayésienne, Effets mixtes, MCMC-SAEM, Classification

**Abstract.** Disease progression models are a promising tool to analyze longitudinal data presenting multiple modalities. Such models can be used to estimate a long-term disease

progression and to reconstruct individual trajectories that can account for the variability between patients but also between modalities. We implemented the multivariate version, combining 14 clinical tests, of a bayesian mixed-effect model (Leaspy) that involves a time reparametrisation to analyze the evolution of cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL). We analyzed data from 395 patients with 2007 visits, recruited at the French National Referral Centre CERVCO, using Leaspy to assess the temporal variability, resulting from different pace of progression and temporal offset, along with the spatial variability, that takes the form of a modification in the sequence of events. Our multivariate analysis allowed us to model the disease progression as a 14-dimensional vector, and thus account for the relationships between the dependent variables, that showed variations in the starting deterioration timepoint. We then implemented a GMM clustering algorithm for the individual parameters to identify underlying subtypes for the disease. We identified a subgroup of patients with early onset and fast progression, showing motor symptoms and neurological deficits earlier, and a group with late onset and slow progression, showing cognitive symptoms earlier. To conclude the multivariate model manages to highlight the differences in the evolution of clinical scores suggesting a gradual but also heterogenous decline in CADASIL and the profiling revealed two subtypes that imply both the temporal and the spatial variability.

**Keywords.** Disease Progression Model, Multivariate, Bayesian, Mixed-effects, MCMC-SAEM, Classification

# 1 Introduction

Les modèles à effets mixtes sont prometteurs pour analyser des données répétées représentant plusieurs modalités et sont souvent utilisés dans les études épidémiologiques. De tels modèles peuvent être utilisés pour estimer une progression à long terme d'une maladie et reconstruire des trajectoires individuelles qui tiennent compte de la variabilité entre les patients (effets aléatoires), mais aussi entre les modalités (effets fixes). Comprendre comment la maladie progresse et quelle est la variabilité attendue entre les individus est essentiel. L'hétérogénéité de la maladie est souvent influencée par certaines covariables, mais parfois elle est plus complexe comme elle est structurée autour du sous-type auquel l'individu appartient. Si les modéles classiques font l'hypoothése d'une évolution homogène des patients, ceci n'est pas forcément réaliste car il existe souvent des sous-groupes des patients évoluant de façon différente. La découverte de clusters inconnus d'individus est un défi complexe qui s'ajoute à l'estimation des paramètres du modèle. Dans ce cadre les modèles de progression de maladie (*Disease Progression Models*) sont un outil émergent qui reconstruit les chronologies des maladies chroniques à long terme, fournissant ainsi un aperçu unique des processus pathologiques et de leurs mécanismes sous-jacents (Young et al. (2024)). Ils sont interprétables, facilitant ainsi la compréhension des maladies tout en permettant la classification, la prédiction et la stratification.

Dans cette étude nous avons utilisé un modèle de progression des maladies, non-linéaire

à effets mixtes qui passe par une reparamétrisation du temps (Leaspy) comme proposé pas Schirrati et al. (2017). Nous avons implémenté une version logistique multivariée du modéle sur les donnés de la maladie CADASIL en évaluant l'intensité de 14 scores cliniques dans le but de construire une cartographie du cours de la maladie. En utilisant les paramètres individuelles, nous avons identifié deux sous-groupes présentant des trajectoires de maladie différentes.

## 2 Modèle de progression (Leaspy)

Nous considérons des données longitudinales, c'est-à-dire des patients avec des mesures répétées dans le temps correspondant à 14 scores cliniques mesurant différents aspects de la maladie. Le modèle de progression utilisé (Leaspy), a déjà fait ses preuves pour décrire la progression des maladies de façon multivarié, et a déjà été implémenté pour diverses maladies neurodégénératives (Ortholand et al. (2023), Koval et al. (2021,2022)). Nous supposons un ensemble de données longitudinales $(y_{ijk})_{1\leq i\leq n, 1\leq j\leq N_i, 1\leq k\leq d}$ qui decrivent les valeurs pour chaque patient $i$ dans la visite $j$ de la caractéristique $k$. Le nombre total de visites $N_i$ peut être différent pour chaque patient $i$. Les données $y_{ijk}$ sont supposées être des points sur une variété riemannienne $\mathcal{M}$. Les paramètres de population (effets fixes) décrivent la trajectoire moyenne comme une géodésique $\gamma_0$ sur la $\mathcal{M}$ avec $\gamma_0(t_0) = \mathbf{p}$ et $\dot{\gamma_0}(t_0) = \mathbf{v}$. Ici nous utilisons la forme logistique du modèle :

$$y_{ijk} = (1 + (\frac{1}{p_k} - 1)exp(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1 - p_k)}))^{-1} + \epsilon_{ijk} \tag{1}$$

avec une erreur résiduelle $\epsilon_{ij} \sim \mathcal{N}(\mathbf{O}_d, \sigma^2\mathbf{I}_d)$. Pour simplifier, notre modèle construit la trajectoire moyenne de chaque score sous la forme d'une courbe logistique décrite par les paramètres $p_0$, $v_0$ et $t_0$, où $p_0$ et $v_0$ sont la position et la vitesse (dérivée de la courbe) au temps $t_0$, le point médian de la logistique. La reparamétrisation du temps individuel prend le forme suivante : $\psi_i(t) = \alpha_i(t-t_0-\tau_i)+t_0$ où $\tau_i$ est le décalage temporel et $\alpha_i = e^{\xi_i}$ est le facteur d'accélération. Le profil temporel des patients est décrit par deux paramètres : le décalage temporel $\tau_i$ qui correspond à l'âge estimé du début de la maladie, mesuré en années, et le taux de progression $\xi_i$ qui indique si la progression globale est accélérée (valeurs positives) ou ralentie (valeurs négatives). Le profil spatial est défini par les paramètres d'espacement intermarqueurs $(\omega_i)_i$ qui tiennent compte de l'ordre différent des événements au sein de notre population. Pour chaque patient, nous avons un $\omega$ pour chaque score clinique inclus dans le modèle avec des valeurs négatives indiquant qu'un score spécifique commence à se détériorer plus tôt pour un patient spécifique que pour la population moyenne. Néanmoins pour une interprétabilité accrue, le modèle utilise une analyse en composantes indépendantes avec $N_s$ sources indépendantes, ce qui conduit à des paramètres de décalage spatial $\mathbf{w}i = As_i$ tels que les colonnes $A_l = \sum_{k=1}^{d-1} \beta_{lk}B_k$ sont une combinaison linéaire d'une base orthonormale $(B_k)_{1\leq k\leq d-1}$ du sous-espace orthogonal à $Span(\mathbf{v})$.

Le modèle statistique hiérarchique suppose que les paramètres de population $z_{pop} =$

$(\mathbf{g}, \mathbf{v}, t_0, A)$ et les paramètres individuels $z_i = (\xi_i, \tau_i, \mathbf{w}_i)$ sont latents et suivent des distributions gaussiennes :

$$\begin{cases} \xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \\ \tau_i \sim \mathcal{N}(0, \sigma_\tau^2) \\ s_i \sim \mathcal{N}(\mathbf{0}_{N_s}, \mathbf{I}_{N_s}) \end{cases} \quad (2) \qquad \begin{cases} g_k = \frac{1}{p_k} - 1 \quad et \quad \mathbf{g} \sim \mathcal{N}(\bar{\mathbf{g}}, \sigma_g^2 \mathbf{I}_d) \\ v_k = e^{\tilde{v}_k} \quad et \quad \tilde{\mathbf{v}} \sim \mathcal{N}(\bar{\mathbf{v}}, \sigma_v^2 \mathbf{I}_d) \\ t_0 \sim \mathcal{N}(\bar{t_0}, \sigma_t^2) \\ \beta_{lk} \sim \mathcal{N}(\bar{\beta}_{lk}, \sigma_\beta^2) \end{cases} \quad (3)$$

Enfin les paramètres du modèle statistique sont $\theta = (\sigma_\xi, \sigma_\tau, \bar{\mathbf{g}}, \bar{\mathbf{v}}, \bar{t_0}, \bar{\beta}_{lk})$ tandis que $\sigma_g$, $\sigma_t$, $\sigma_v$, $\sigma_\beta$ sont fixes.

# 3 Application sur des données réelles

## 3.1 CADASIL

L'artériopathie cérébrale autosomique dominante à l'origine d'infarctus sous-corticaux et d'une leucoencéphalopathie (CADASIL) est la maladie héréditaire la plus courante touchant les petits vaisseaux cérébraux (SVD) et elle est causée par des mutations du gène NOTCH3 (Rutten et al. (2020)). Les patients atteints de CADASIL présentent un large éventail de symptômes telles que des crises de migraines avec aura, des accidents vasculaires cérébraux, des troubles du comportement, une incapacité motrice et des troubles cognitifs de la dysfonction exécutive jusqu'à une démence sévère (Chabriat et al. (2009)). Les premiers signes de déclin cognitif sont observés principalement au niveau des fonctions exécutives et de la mémoire à court terme et des altérations significatives apparaissent progressivement dans tous les domaines cognitifs (Buffon et al. (2006), Amberla et al. (2004), Peters et al (2005), Brookes et al. (2016)). CADASIL se manifeste cliniquement entre 40 et 70 ans et conduit à une démence vasculaire sous-corticale pour la grande majorité des patients (Viitanen et Kalimo (2000)). Les manifestations cliniques présentées tout au long de la maladie varient en intensité au fil du temps et chez différents groupes de patients (Brice et al. (2020,2022)). Malheureseument, malgré ces informations sur l'histoire naturelle du CADASIL la cinétique exacte du déclin tout au long de la maladie reste encore trés mal connues. Dans notre étude, nous avons implémenté un modèle Leaspy multivarié en prenant en compte 14 scores cliniques et d'évaluations afin de capturer l'ensemble du spectre du déclin causé par la maladie.

## 3.2 Données

Nous avons analysé les données de 395 patients effectuant au total 2007 visites avec une médiane de 4 visites par patient (IQR : 3-7), recrutés au Centre National de Référence CERVCO en France. La durée de suivi a varié de 6 mois à 19 ans, avec une moyenne de 7.5 ans (écart-type : 4.7 ans), et l'âge d'inclusion de 25 à 80 ans, (moyenne de 52,2 ans et écart-type : 11.7). Ces patients avaient complété au moins deux visites, le temps de visite n'était pas manquant et ils présentaient au moins l'un des 14 scores déterminés par consensus entre

les neurologues (N.A., D.H. et H.C.) et les neuropsychologues (A.J., S.R., C.M. et C.R.) du centre de référence, obtenus lors du suivi.

Les scores sélectionnés peuvent nous fournir une description du declin global qui est observé pendant toute la durée de la maladie. L'efficacité cognitive a été évaluée à l'aide de l'échelle de Mattis (MDRS : Mattis Dementia Rating Scale) et de la sous-échelle VADAS-Cog (VADAS-Cog : Vascular Dementia Assessment Scale cognitive subscale). Le score d'Initiation/Persévérance (*MDRSinitiation*) du MDRS a été pris en compte. Trois des composants de la VADAS-Cog ont été inclus dans notre analyse : le Digit Cancellation Test (*DigitCancel*), le Symbol Digit Test(*SymbolDigit*) et le Backward Digit Span (*BackwardDigit*). Nous avons également pris en compte les 3 scores obtenus du Trail Making Test (TMT) évaluant la vitesse cognitive et la flexibilité mentale, le temps pour la partie A du TMT (*TMTAT*), le temps et les erreurs pour la partie B du TMT (*TMTBT*, *TMTBE*). Les performances de la mémoire ont été analysées à l'aide de 3 scores de l'épreuve de rappel libre/ rappel indicé du Grober et Buschke (GB) : le rappel libre total (*GBfree*), l'indice de sensibilité au repérage (*GBcueing*) et le rappel total indicé (*GBdelayed*). La sévérité du handicap a été évaluée à l'aide de l'échelle de Rankin modifiée (*Rankin*) et l'indépendance dans les activités quotidiennes par l'indice de Barthel (*Barthel*). Les déficits neurologiques focaux ont été aussi inclus dans notre analyse, représentés par le score NIH pour les AVC (*NIHSS*). Enfin, la qualité de vie a été évaluée à l'aide de l'échelle visuelle analogique de l'EuroQol à 3 niveaux (*EQVAS*). Ces 14 scores ont été transformés et normalisés pour obtenir une évolution ascendante de 0 à 1, répondant aux exigences de notre modèle.

## 3.3 Résultats

### 3.3.1 Évolution du CADASIL

Basé sur le modèle Leaspy le plus performant pour l'évolution du CADASIL, nous observons qu'il existe trois groupes de scores qui se mofdifient à des stades différents de la maladie (Figure 1). Le premier groupe de scores, dont l'évolution commence à un stade précoce de la maladie, comprend des mesures de dysfonctionnement exécutif comme les sous-scores VADAS-Cog, Symbol Digit Test (*SymbolDigit)*, Backward Digit Span (*BackwardDigit*) et Cancellation Task (*DigitCancel*), TMT B temps (*TMTBT*) et aussi certains domaines de performance de la mémoire comme le GB Total Free Recall (*GBfree*). Aux stades intermédiaires de la maladie, nous pouvons voir que la qualité de vie EQ VAS (*EQVAS*), la sévérité du handicap mesuré par l'échelle de Rankin modifiée (*Rankin*), et le temps du TMT A (*TMTAT*) commencent également à montrer une évolution. Enfin, à un stade plus avancé de la maladie, les caractéristiques qui commencent à se détériorer correspondent aux divers domains comme l'éfficacité cognitive représentée par l'Initiation du Mattis DRS (*MDRSinitiation*) et les erreurs du TMT B (*TMTBE*), la performance de mémoire mesuré par le GB Index of Sensitivity to Cueing (*GBcueing*) et le GB Delayed Total Recall (*GBdelayed*) et assi les activités quotidiennes(*Barthel*) et les AVC (*NIHSS*).

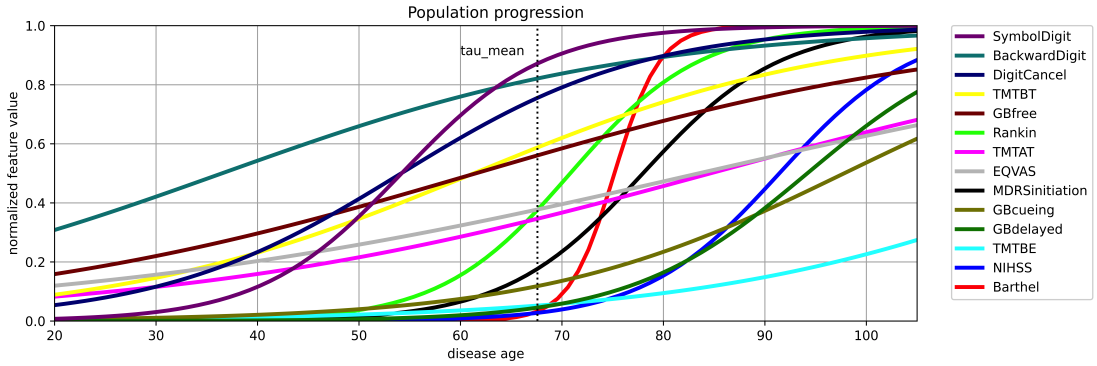



FIGURE 1 – Évolution du CADASIL décrite par la progression longitudinale moyenne de 14 scores cliniques.

### 3.3.2 Sous-groupes de patients avec des trajectoires similaires

Pour identifier les sous-groupes de patients avec des trajectoires similaires, nous avons appliqué un modèle de mélange gaussien (GMM) (Zhou (2021)). À cette fin, nous avons pris en compte les paramètres individuels $(\xi_i, \tau_i, s_i)$ et nous avons fixé le nombre de clusters à 2 qui optimisent le critère d'information d'Akaike et le critère d'information Bayésien (AIC et BIC).

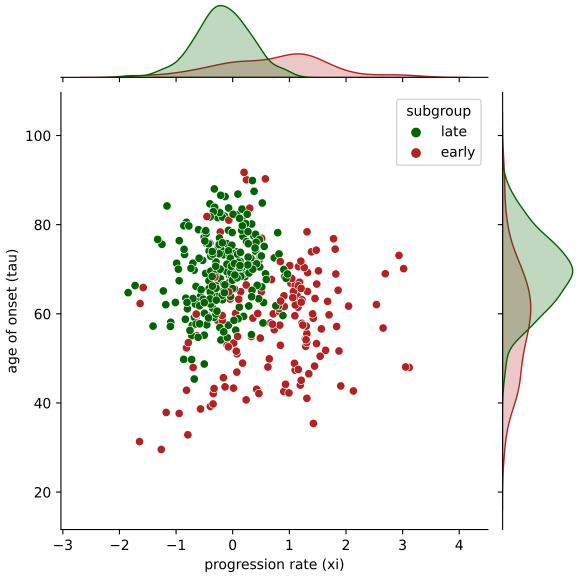



FIGURE 2 – Répartition en sous-groupes en fonction des paramètres temporels individuels ($\tau$ et $\xi$).

Leur répartition est représentée dans la figure 2 selon les paramètres temporels $\tau$ et $\xi$. Sur les axes supérieur et droit, on peut également voir les densités spécifiques par sous-groupe des paramètres individuels. Le premier sous-groupe, précoce (ou rapide), comprend 37% des patients avec un tau moyen de 58,27 (écart type 12,20) et un xi moyen de 0,73 (0,95). Les 63% restants des patients appartiennent au deuxième sous-groupe, tardif (et lent), et présentent un tau moyen de 69,33 (8,20) et un xi de -0,18 (0,49). Une fois que les clusters ont été formés, afin de décrire le profil des patients attribués à chaque sous-groupe, nous avons réalisé un test U de Mann-Whitney pour identifier les éventuelles différences dans les distributions des paramètres $\tau_i$ , $\xi_i$ et $\omega_i$ entre les sous-groupes. Après avoir examiné les distributions des $\omega$, nous constatons que le sous-groupe précoce (ou rapide) présente des symptômes d'incapacité motrice plus tôt

que l'autre sous-groupe tardif (et lent), car il a des omegas plus petits pour les indices de Barthel et de Rankin. Dans ce sous-groupe, nous observons également un début plus précoce des accidents vasculaires cérébraux, comme le montrent les distributions de l'indice NIHSS. D'autre part, les patients du groupe tardif (et lent) présentent un début plus précoce des symptômes cognitifs, comme le montrent les différences dans les distributions de l'Initiation du MDRS, des scores TMT, du GB Total Free Recall et des sous-scores VADAS-Cog. Les distributions de la qualité de vie EQ VAS et de GB Delayed Total Recall et l'Index of Sensitivity to Cueing sont similaires pour les deux sous-groupes. Toutes les distributions $\omega$ selon les sous-groupes sont montrées dans la figure 3.

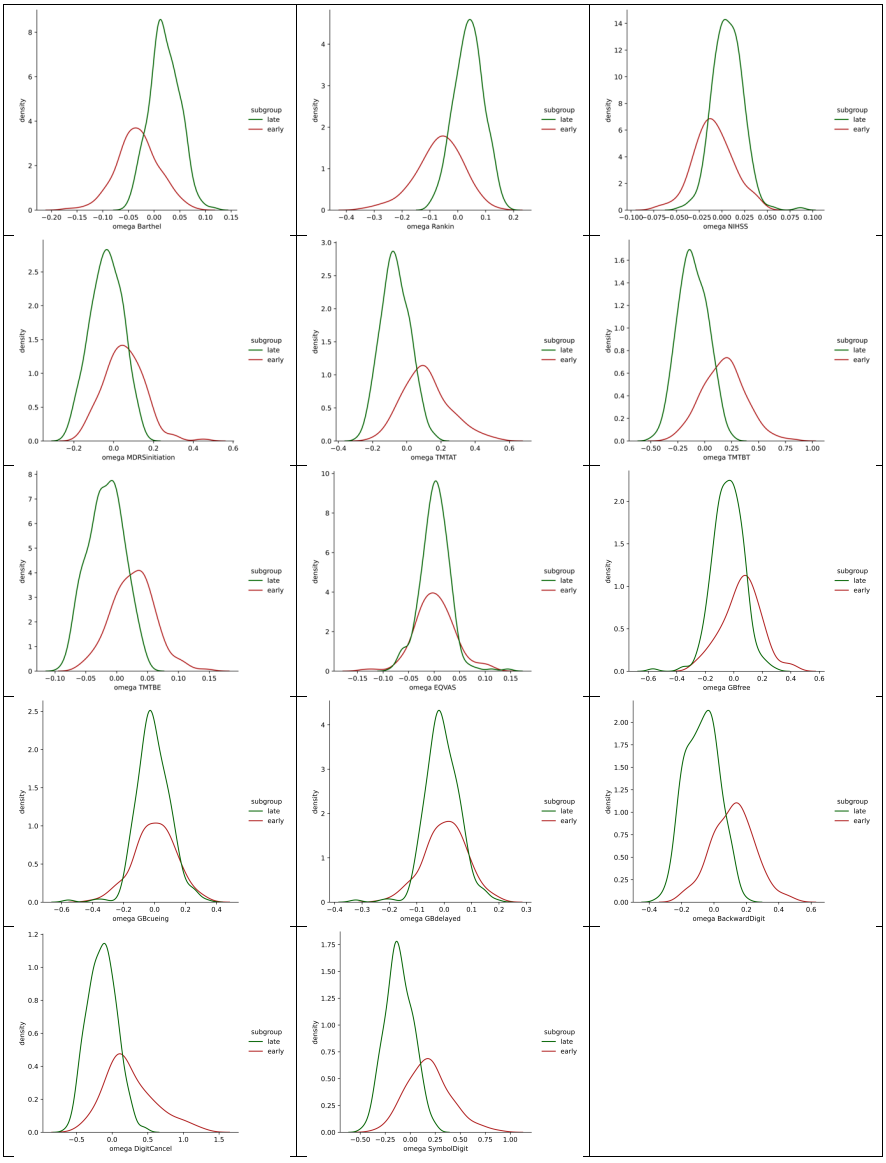



FIGURE 3 – Distributions spécifiques au sous-groupe des paramètres spatiaux individuels ($\omega$).

# 4 Discussion

Cette étude permet de mieux comprendre l'histoire naturelle de la maladie et d'explorer individuellement les patients pour en déterminer des groupes d'évolution différente. Bien que la variété en termes d'âges, de moments de début et de progression puisse poser des difficultés pour construire un modèle fiable, la reparamétrisation de l'âge nous permet de construire une trajectoire de population cohérente avec le cours de la maladie et de comparer les individus se trouvant au même stade de la maladie plutôt que d'avoir le même âge biologique. Un atout particulièrement fort est que nous sommes en mesure d'examiner le début de la maladie en tant que paramètre individuel $\tau$ estimé par notre modèle, alors que ce début est difficilement défini cliniquement pour les patients attaints du CADASIL. Dans la plupart des études, le début était généralement considéré comme le moment où le premier accident vasculaire cérébral se produit (Amberla et al. (2004), Peters et al. (2005), Chabriat et al. (2009)). La migraine avec aura est également mentionnée comme un symptôme inaugural majeur du CADASIL et peut survenir des années avant le premier accident vasculaire cérébral (Guey et al. (2016)).

Le résultat le plus prometteur de notre étude est qu'elle a révélé deux phénotypes de CADASIL, plus complexes que ce qui était initialement soupçonné, avec des trajectoires qui diffèrent non seulement en termes d'évolution temporelle, mais aussi en ce qui concerne la séquence d'événements. Le profilage effectué en considérant simultanément tous les aspects de l'évolution spatiotemporelle individuelle nous a permis de conclure que les patients présentant une forme plus agressive de la maladie développent également des symptômes précoces d'incapacité motrice et d'accidents vasculaires cérébraux, tandis que chez les patients présentant une forme légère, les symptômes cognitifs apparaissent en premier.

Il y a également plusieurs limites dans cette étude qui laissent place à des investigations supplémentaires. Les clusters individuels ont été formés a posteriori en utilisant les paramètres individuels estimés par notre modèle. Une procédure de classification intégrée à l'étape d'estimation du modèle, en supposant un modèle de mélange peut nous fournir une description plus précise et non-biasé de la trajectoire de la maladie ainsi que des différents phénotypes sous-jacents.

# 5 Conclusion

Avec notre modèle de progression multivarié, nous avons pu obtenir des résultats fiables et prometteurs sur la cartographie de l'évolution du CADASIL et sur ses variations au sein de différents groupes de patients. Nous considérons cela comme une première étape, car le modèle, dans son état actuel, offre encore de nombreuses possibilités pour approfondir la compréhension de la cinétique de la maladie. Les sous-groupes suggérés ici illustrent que les phénotypes sous-jacents de CADASIL sont potentiellement plus complexes et implique le profil temporel et aussi spatial de patients. Cette perception nous offre des perspectives d'amélioration afin de profiler de manière plus efficace les différents sous-groupes.

# Bibliographie


Amberla, K., Wäljas, M., Tuominen, S., Almkvist, O., Pöyhönen, M., Tuisku, S., Kalimo, H., et Viitanen, M. (2004), Insidious cognitive decline in CADASIL. *Stroke*, 35(7), pp. 1598-1602.

Brice, S., Jabouley, A., Reyes, S., Machado, C., Rogan, C., Dias-Gastellier, N., Chabriat, H., et du Montcel, S. T. (2020), Modeling the Cognitive Trajectory in CADASIL. *Journal of Alzheimer's disease : JAD,* 77(1), pp. 291-300.

Brice, S., Reyes, S., Jabouley, A., Machado, C., Rogan, C., Gastellier, N., Alili, N., Guey, S., Jouvent, E., Hervé, D., Tezenas du Montcel, S., et Chabriat, H. (2022), Trajectory Pattern of Cognitive Decline in Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy. *Neurology*, 99(10), pp. e1019-e1031.

Brookes, R. L., Hollocks, M. J., Tan, R. Y., Morris, R. G., et Markus, H. S. (2016), Brief Screening of Vascular Cognitive Impairment in Patients With Cerebral Autosomal-Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy Without Dementia. *Stroke*, 47(10), pp. 2482-2487.

Buffon, F., Porcher, R., Hernandez, K., Kurtz, A., Pointeau, S., Vahedi, K., Bousser, M. G., et Chabriat, H. (2006), Cognitive profile in CADASIL. *Journal of neurology, neurosurgery, and psychiatry*, 77(2), pp. 175-180.

Chabriat, H., Joutel, A., Dichgans, M., Tournier-Lasserve, E. et Bousser, M.G. (2009), Cadasil. *The Lancet. Neurology*, 8(7), pp. 643-653.

Guey, S., Mawet, J., Hervé, D., Duering, M., Godin, O., Jouvent, E., Opherk, C., Alili, N., Dichgans, M., et Chabriat, H. (2016), Prevalence and characteristics of migraine in CADASIL. *Cephalalgia : an international journal of headache*, 36(11), pp. 1038-1047.

Koval I., Bône A., Louis M.,Lartigue T., Bottani S., Marcoux A., Samper-González J., Burgos N., Charlier B., Bertrand A., Epelbaum S., Colliot O., Allassonnière S. et Durrleman S. (2021), AD Course Map charts. *Scientific Reports - Nature*, 11(1), 8020

Koval, I., Dighiero-Brecht, T., Tobin, A. J., Tabrizi, S. J., Scahill, R. I., Tezenas du Montcel, S., Durrleman, S., et Durr, A. (2022), Forecasting individual progression trajectories in Huntington disease enables more powered clinical trials. *Scientific reports - Nature*, 12(1), 18928.

Ortholand J., Pradat P.F., Tezenas du Montcel S., et Durrleman S. (2023) Interaction of sex and onset site on the disease trajectory of amyotrophic lateral sclerosis. *Journal of Neurology*, 270(12), pp. 5903-5912.

Peters, N., Opherk, C., Danek, A., Ballard, C., Herzog, J., et Dichgans, M. (2005), The pattern of cognitive performance in CADASIL : a monogenic condition leading to subcortical ischemic vascular dementia. *The American journal of psychiatry*, 162(11), pp. 2078-2085.

Rutten, J. W., Hack, R. J., Duering, M., Gravesteijn, G., Dauwerse, J. G., Overzier, M., van den Akker, E. B., Slagboom, E., Holstege, H., Nho, K., Saykin, A., Dichgans, M., Malik, R., et Lesnik Oberstein, S. A. J. (2020), Broad phenotype of cysteine-altering NOTCH3 variants in UK Biobank : CADASIL to nonpenetrance. *Neurology*, 95(13), pp. e1835-e1843.

Schirratti J.B., Allassonniére S., Colliot O., et Durrleman S. (2017), A Bayesian Mixed- Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations. *Journal of Machine Learning Research*, 18, pp. 1-33.

Viitanen, M., et Kalimo, H. (2000), CADASIL : hereditary arteriopathy leading to multiple brain infarcts and dementia. *Annals of the New York Academy of Sciences*, 903, pp. 273-284.

Young, A. L., Oxtoby, N. P., Garbarino, S., Fox, N. C., Barkhof, F., Schott, J. M., et Alexander, D. C. (2024), Data-driven modelling of neurodegenerative disease progression : thinking outside the black box. *Nature reviews.* Neuroscience, 25(2), pp. 111-130.

Zhou Z.H. (2021), *Machine Learning*. Springer, Singapore. pp. 222-227

# A Regression Model on Quantiles Extracted from Scans of Asthmatic Patients

Marie-Felicia Beclin [1] & Pierre Lafaye de Micheaux [2] & Nicolas Molinari[3]

[1] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, France, marie-felicia.beclin@umontpellier.fr*
[2] *UNSW Sydney lafaye@unsw.edu.au*
[3] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, France, nicolas.molinari@inserm.fr*

**Résumé.** Nous nous intéressons à l'évaluation de l'efficacité du Benralizumab, un médicament utilisé pour traiter l'asthme, en utilisant des scans tomographiques capturés pendant l'expiration et l'inspiration avant et après un an de traitement. L'hypothèse médicale de travail postule que les patients dont l'état s'est amélioré présenteront des images de scanners thoraciques en expiration améliorées après le traitement. C'est-à-dire que le patient expire mieux et donc son poumon se vide plus en expiration. Cela se manifeste par des valeurs d'unité Hounsfield plus élevées. Il y a alors un déplacement vers la droite dans l'histogramme construit à partir de l'image post-traitement par rapport à celui pré-traitement.

Irpino et Verde[1] ont adopté la méthode classique de la régression linéaire de manière à pouvoir l'appliquer aux fonctions quantiles plutôt qu'aux observations réelles. Nous généralisons leur approche et obtenons les lois des estimateurs des paramètres du modèle via une approche de maximum de vraisemblance. À partir de l'espace $S_Q$ des fonctions quantiles, nous définissons des polynômes de quantiles. Nous nous penchons ensuite sur le cas particulier linéaire. Nous définissons alors explicitement les estimateurs pas maximum de vraisemblance.

Le modèle a été implémenté en Python et appliqué à un ensemble de données réelles de 40 patients traités par Benralizumab.

L'approche décrite ci-dessus présente certaines limites, notamment la perte d'informations spatiales et l'hypothèse de relations linéaires entre les distributions de voxels. Des recherches supplémentaires sont nécessaires pour développer une méthode de régression plus générale, telle que celle proposée par Chen[2] et Ghodrati et Panaretos[3]. Cependant, notre approche a l'avantage d'être simple, facile à utiliser et comprise par les praticiens. Le recalage des images en inspiration et en expiration[4] permet une correspondance voxel par voxel. Nous pouvons alors généraliser cette approche précédente. Des recherches en cours visent à prédire des histogrammes 2D post-traitement à partir de scanners pendant l'inspiration et l'expiration après enregistrement, ainsi que des histogrammes pré-traitement correspondants, tout en incluant des covariables scalaires.

**Mots-clés.** Régression de distribution sur distribution, Fonctions quantiles, Biomarqueur dérivé de l'imagerie, Prédiction de traitement, Histogrammes.

**Abstract.** We are interested in evaluating the efficacy of Benralizumab, a medication to treat asthma, by using tomography scans captured during expiration and inspiration

before and after one year of treatment. The medical working hypothesis posits that patients with improved conditions will exhibit enhanced expiration scans after treatment, which is manifested by higher Hounsfield unit values. This results in a shift to the right in the histogram built from post-treatment image compared to the pre-treatment one.

Irpino and Verde[1] mimicked the classical linear regression method so that it can be applied to quantile functions instead of real-valued observations. We generalize their approach and obtain confidence intervals and laws of the estimators in the model via a maximum likelihood approach. From the space $S_Q$ of quantile functions, we define quantile polynomials. We then focus on the specific linear case. We explicitly define the maximum likelihood estimators.

The model was implemented in a Python code and applied to a real data set of 40 patients treated by Benarlizumab.

The approach described above has some limitations, including the loss of spatial information and the assumption of linear relationships between voxel distributions. Further investigation is needed to develop a more general distribution-on-distribution regression method, such as the works by Chen[2] and Ghodrati and Panaretos[3]. But our approach has the advantage of being simple, easy to use and understood by practitioners. The registration of images in inspiration and expiration[4] allows for a voxel-to-voxel correspondence. We can then generalize our previous approach. Ongoing research aims to predict post-processing 2D histograms from CT scans during inspiration and expiration after recording, as well as corresponding pre-processing histograms, all while including scalar covariates.

**Keywords.** Distribution on Distribution Regression, Quantiles Functions, Imaging-derived biomarker, Treatment prediction, Histograms.

# 1 Des images aux Quantile

## 1.1 Données cliniques

Tous les patients de l'étude ont été traités par Benralizumab sur une période de 48 semaines. Le médicament a été administré par voie sous-cutanée à une dose de 30 mg par injection. Les données récoltées sont diverses : données cliniques ; mesure de l'examen fonctionnel respiratoire et enfin des images de scanner thoracique 3D en inspiration et en expiration. Une image scanner se compose d'environ 600 images 2-D de dimensions $512 \times 512$. Les données brutes sont stockées dans des fichiers au format DICOM et, lors du prétraitement, chaque scan est converti en un tableau numérique de dimensions $600 \times 512 \times 512$ pixels. Les valeurs des données des voxels sont exprimées en valeurs de Hounsfield (HU). Les poumons sont ensuite segmentés en appliquant un algorithme de segmentation par seuillage[4] ou par réseau neuronal [5]. Pour chaque patient $i$, $N_i$ voxels sont sélectionnés par segmentation en fonction de la taille du poumon du patient. La valeur d'unité Hounsfield la plus basse possible, -1024 HU, correspond à de l'air. Nous utilisons les histogrammes des valeurs d'unité Hounsfield avant et après le traitement comme méthode pour identifier les répondeurs cliniques au Benralizumab.

Cependant, il est important de noter que par cette méthode, nous sacrifions des informations sur la distribution spatiale des voxels.

## 1.2 Les fonctions quantiles

Pour une image 3D en niveaux de gris , on passe à l'histogramme des valeurs.

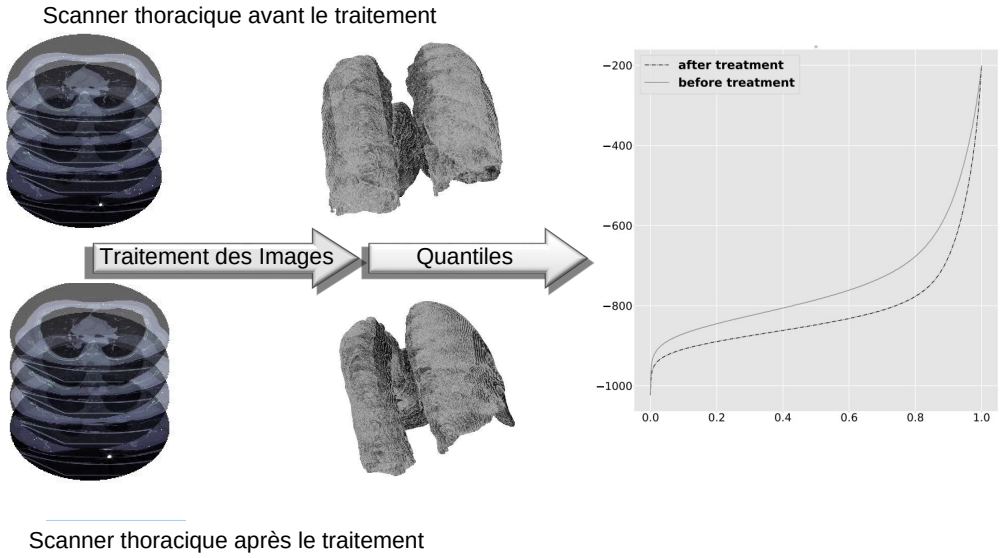



Figure 1: Illustration du processus d'extraction des données quantiles à partir des images de scanners thoraciques

La fonction quantile $q^X$ associée à une variable aléatoire est définie pour tout $p \in ]0,1[$ :

$$q^X(p) \quad = \quad \inf_{t\in\mathbb{R}}\{t, F_X(t) \geq p\} \text{ avec } F^X \text{ la fonction de répartition de } X. \tag{1}$$

Nous noterons $x_q = \int_0^1 q^X(p)dp$. Ainsi, nous introduisons l'espace des fonctions quantiles, de carré intégrable, doté de la distance suivante : $d(q^X, q^Y) = \sqrt{\int_0^1 (q^X(t) - q^Y(t))^2 dt}$. Les données sur lesquelles nous appliquons notre nouveau modèle de régression sont donc $n$ paires de fonctions quantiles $(q_1^X, q_1^Y), \ldots, (q_n^X, q_n^Y)$ extraites des scans pris sur $n$ sujets avant et après le traitement.

## 1.3 Le modèle

**Polynômes de Quantiles** Avant d'introduire notre modèle statistique, nous devons introduire les quantiles polynomiaux. Nous dénoterons l'espace des fonctions quantiles par $\mathrm{S}_q$. Pour réduire cet espace et obtenir un espace de travail plus "confortable", nous travaillerons

sur les quantiles polynomiaux du quantile $\mathcal{N}(0,1)$, notée $q$. Pour cela, il faut se restreindre aux polynômes $P$ tels que $P \circ q$ est une fonction quantile.

Supposons que nous disposions d'une loi de probabilité $\mathcal{QP}(\boldsymbol{\theta}, L)$ sur cet espace. Nous pouvons alors définir $E_1, \dots E_n$ les éléments aléatoires quantiles suivant $\mathcal{QP}(\boldsymbol{\theta}, L)$. On note $E_k = \sum_{k=1}^{L} A_{i,k} q^i$ .

Soit $\forall i \in \{1, 2, \dots, n\}$ , $Q_i^Y : \Omega \to \mathrm{S}_q$ qui satisfait :

$$Q_i^Y(p) = \alpha + b_1 x_{q_i} + b_2 q_i^{cx}(p) + E_i(p), \text{ avec } b_2 > 0 \tag{2}$$

Puisque $q^{cx}$ est une fonction quantile de la loi $N(0, \sigma^2)$ $q^{cx} = \sigma q$. Donc, $\alpha + b_1 x_{q_i} + b_2 q_i^{cx}(p) + E_i(p)$ peut être écrit comme $S \circ q$, avec $S$ un polynôme.

$$Q_i^Y \quad = \quad \alpha + b_1 x_{q_i} + b_2 \sigma_i q + \sum_{k=0}^{L} A_{i,k} q^i = \sum_{k=0}^{L} B_{i,k} q^k \tag{3}$$

À l'instar d'une régression linéaire, nous montrons alors que pour tout $i$, $Q_i^Y \sim \mathcal{QP}(\boldsymbol{\theta_i})$. Ainsi, en réussissant à définir la loi $\mathcal{QP}(\boldsymbol{\theta_i})$, nous pourrons avoir une expression explicite de la vraisemblance.

**Le cas** $L = 1$ Pour le cas $L = 1$, nous pouvons expliciter une loi sur l'espace des polynômes de quantile et donc créer un modèle avec des solutions explicites.

$$S_{QP_1} := \{r = m + \sigma q\} = \left\{q : ]0,1[ \to \mathbb{R}; \quad \exists m \in \mathbb{R}, \exists s^2 > 0 \text{ tel que } q \text{ quantile de } \mathcal{N}\left(m, s^2\right)\right\}$$

Définissons $Q$ un élément aléatoire sur cet espace de quantile. Soit $\mu \in \mathbb{R}$ et $\sigma^2, \beta > 0$,

$$\begin{array}{rccl} Q := Q_{\mu,\sigma^2,\beta} : \Omega & \to & S_{q_{\mathcal{N}}} & \\ \omega & \mapsto & q_X := Q(\omega), \text{ quantile de X } & \sim \mathcal{N}(u, v^2), \end{array} \tag{4}$$

avec $u = U(\omega)$ (respectivement $v = V(\omega)$) est une réalisation de la variable $U : \Omega \to \mathbb{R}$ (respectivement $V : \Omega \to \mathbb{R}^+$) et $U \sim \mathcal{N}\left(\mu, \sigma^2\right)$ (respectivement $V \sim \mathcal{E}xp\left(\beta^{-1}\right)$) avec $U$ et $V$ indépendants.

$Q$ est une variable aléatoire dont on peut définir une fonction de densité. On note $Q\mathcal{N}(0, \sigma^2, \beta^{-1})$ la loi de $Q$.

Soient $q_1^X, q_2^X, \dots, q_n^X$ des fonctions quantiles déterministes prédictives de $S_{QP_1}$ et $\epsilon_1, \dots, \epsilon_n$ des éléments aléatoires quantiles normaux distribués comme $Q\mathcal{N}(0, \sigma^2, \beta^{-1})$.

Soient $Q_1^Y, \dots, Q_n^Y$ i.i.d. d'un élément aléatoire quantile $Q^Y : \Omega \to S_Q$.

Le modèle statistique est défini comme suit :

$$\forall i \in 1, 2, \dots, n, \forall p \in ]0,1[ \quad Q_i^Y(p) = \alpha + b_1 x_{q_i} + b_2 q_i^{cx}(p) + \epsilon_i(p), \tag{5}$$

où les coefficients réels inconnus $\alpha$, $b_1$ et $b_2$ seront estimés via une approche basée sur le maximum de vraisemblance.

La méthode décrite peut être modifiée avec différentes hypothèses, mais la condition importante est que les quantiles ont tous la même forme. Si ce prérequis est respecté, le procédé peut être adapté facilement. Dans le cas d'une fonction de quantile de forme normale, la distance de Wasserstein entre les distributions dépend uniquement de la moyenne et de la variance, ce qui facilite l'explication de la vraisemblance et permet d'obtenir des expressions explicites pour chaque estimateur.

Les estimateurs obtenus sont les suivants :

$$\widehat{b}_1 = \left(\sum_{i=1}^{n} y_{q_i} x_{q_i} - n\bar{y}\bar{x}\right)\left(\sum_{i=1}^{n} x_{q_i}^2 - n\bar{x}^2\right)^{-1}, \qquad \widehat{\beta} = \frac{1}{n-1}\sum_{i=1}^{n}\left(s_i - \min_{j\in\{1,\dots,n\}}\left(\frac{s_j}{v_j}\right)v_i\right),$$

$$\widehat{\alpha} = \bar{y} - \widehat{b}_1\bar{x}, \qquad \widehat{\sigma}^2 = (n-2)^{-1}\sum_{i=1}^{n}(y_{q_i} - \widehat{\alpha} - \widehat{b}_1 x_{q_i})^2,$$

$$\widehat{b}_2 = \frac{n}{n-1}\min_{i\in\{1,\dots,n\}}\left(\frac{s_i}{v_i}\right) - \frac{\bar{s}}{(n-1)\bar{v}} \text{ avec } \bar{s} = n^{-1}\sum_{i=1}^{n} s_i,$$

**Qualité des estimations** Nos résultats nous permettent d'expliciter l'intervalle de confiance pour chaque estimateur, ainsi qu'une distance de Cook et une quantité $PseudoR^2$.

# 2 Travail en cours

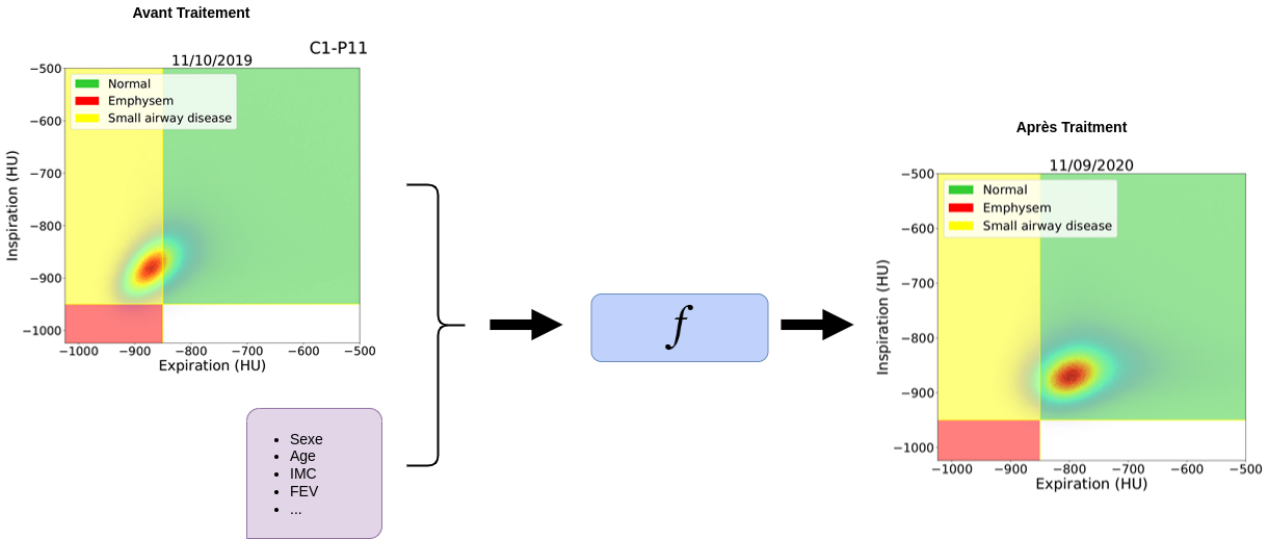



Figure 2: Parametric response map avant et après traitement

Le modèle précédent est utilisé sur les histogrammes de scanner CT en expiration, ce qui est pertinent car l'asthme est une maladie de l'expiration. Cependant, l'interaction entre

l'inspiration et l'expiration n'est pas exploitée. C'est pourquoi nous procédons à un recalage intra-patient de l'image en expiration sur celle en inspiration pour obtenir une correspondance voxel à voxel. Nous utilisons un recalage par B-spline[7].

Ainsi, nous pouvons calculer un histogramme 2D (Figure 2 [1]). Cette "Parametric response map" est introduite par Galbàn[8]. Cela crée un ensemble de données unique. Chaque patient est représenté par une distribution de support bidimensionnel et par des variables cliniques. La question est donc de savoir comment trouver un estimateur basé sur une distribution 2D et une covariable dans l'espace euclidien.

# 3 Bibliographie


1. Irpino, Verde (2015). **Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance.** *Advances in Data Analysis and Classification*, v.9, n.1,p.81–106.

2. Chen,Lin, Müller (2023). **Wasserstein Regression.** *Journal of the American Statistical Association* v.118, n.542, p.869–882.

3. Ghodrati L., Panaretos V. (2022) **Distribution-on-distribution regression via optimal transport maps.** *Biometrika*, v.109, p.957–974.

4. Coselmon, M.M., Balter, J.M., McShan, D.L. Kessler, M.L. (2004), **Mutual information based CT registration of the lung at exhale and inhale breathing states using thin-plate splines.** *Medical Physics* , 31: 2942-2948.

5. Heuberger, J., Geissbuhler, A., Muller, H. (2005). **Lung CT segmentation for image retrieval using the Insight Toolkit (ITK).** *Medical Imaging and Telemedicine*, 30.

6. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G. (2020). **Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem.** *European Radiology Experimental* , 4(1), 1-13.

7. Galbán CJ, Han MK, Boes JL, Chughtai KA, Meyer CR, Johnson TD, Galbán S, Rehemtulla A, Kazerooni EA, Martinez FJ, Ross BD. **Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression.** *Nature Medecine.* 2012 Nov;18(11):1711-5.


[1]Emphysem = Emphysème; Small airway disease = maladie des petites voies aériennes

# Lead time bias correction in breast cancer screening studies

Marius Robert [1] & Brice Amadeo [2] & Marie Poiseuil [2] & Maja Pohar-Perme [3] & Virginie Rondeau [1]

[1] *Biostatistics team, Inserm U1219, University of Bordeaux, France.*
[2] *EPICENE team, Inserm U1219, University of Bordeaux, France.*
[3] *IBMI, Faculty of Medicine, University of Ljubljana, Slovenia.*

marius.robert@u-bordeaux.fr

**Résumé.** Le dépistage par mammographie joue un rôle crucial dans la détection et le diagnostic du cancer du sein, permettant des traitements précoces, améliorant ainsi les taux de survie. L'évaluation de l'efficacité du dépistage implique de comparer la survie des patients diagnostiqués par dépistage avec ceux diagnostiqués après l'apparition des symptômes. Cependant, cette analyse est sujette à des biais pouvant surestimer les avantages du dépistage. Le biais d'avance au diagnostic se manifeste car le dépistage permet un diagnostic précoce du cancer du sein, entraînant une survie observée plus longue sans réelle amélioration de la date de décès. Deux approches fréquemment utilisées pour remédier à ce biais ont été appliquées à des données réelles pour en faire une comparaison. La première approche repose sur le temps de séjour en phase préclinique, c'est-à-dire la période pendant laquelle la tumeur est détectable par dépistage mais ne provoque pas de symptômes. Une distribution de cette durée peut être obtenue à partir d'un modèle multi-états. La correction proposée par Duffy et al. (2008), soustrait l'espérance de cette distribution à la survie observée. Nous proposons d'améliorer la correction déjà existante en tenant compte de la densité mammaire. L'hypothèse sous-jacente est que des tissus mammaires plus denses réduisent la sensibilité du dépistage, raccourcissant ainsi la durée de la phase préclinique. La deuxième approche, développée par Abrahamsson et al. (2020), est basée sur un modèle de croissance tumorale continue. Ce modèle permet d'estimer la croissance volumique jusqu'à l'âge du patient au diagnostic. L'esprit de cette seconde approche est d'utiliser les estimations d'un modèle pour prolonger la croissance tumorale chez les patients diagnostiqués par dépistage pour estimer le temps jusqu'à un diagnostic symptomatique. Bien que cette dernière méthode soit plus précise dans l'estimation du temps d'avance, elle demeure peu utilisée en raison de sa nouveauté, de sa complexité et du manque de logiciel à disposition. Une standardisation de son utilisation serait pertinente, surtout compte tenu des critiques émises sur la correction de Duffy, jugée excessive.

**Mots-clés.** Cancer du sein, dépistage, biais d'avance, survie, correction.

**Abstract.** Mammography screening plays a crucial role in detecting and diagnosing breast cancer, enabling early treatments and thereby improving survival rates. Evaluating the effectiveness of screening involves comparing the survival of patients diagnosed through screening with those diagnosed after the onset of symptoms. However, this analysis is susceptible to biases that may overestimate the benefits of screening. The lead-time bias arises

because screening allows for the early diagnosis of breast cancer, resulting in a longer observed survival without a true improvement in the date of death. Two commonly used approaches to solve this bias were applied to real data in order to compare them. The first approach is based on the sojourn time in the preclinical phase, i.e., the period during which the tumor is detectable through screening without causing symptoms. A distribution of this duration can be derived from a multi-state model. The correction proposed by Duffy et al. (2008) subtracts the expected value of this distribution from the observed survival. We propose enhancing the existing correction by considering breast density. The underlying hypothesis is that denser breast tissues reduce screening sensitivity, thereby shortening the duration of the preclinical phase. The second approach, developed by Abrahamsson et al. (2020), is based on a continuous tumor growth model. This model allows volume growth to be estimated up to the patient's age at diagnosis. The spirit of this second approach is to use estimates from a model to prolong tumor growth in patients diagnosed by screening to estimate the time until a symptomatic diagnosis. Although this latter method is more precise in estimating lead time, it remains underutilized due to its novelty, complexity and lack of available software. Standardizing its use would be relevant, especially considering criticisms of the Duffy correction, deemed excessive.

**Keywords.** Breast cancer, screening, lead time, survival, bias correction.

# 1 Introduction

## 1.1 Breast cancer screening

Screening is a medical practice that involves searching for disease in an asymptomatic individual. The main objective of screening is to detect the disease at an early stage, before the onset of symptoms, which can significantly improve treatment outcomes and therefore patients' survival. Organized screening programs implemented in developed countries provide for mammograms at regular intervals for women aged over 50 years, generally every 2 years. In screening programs, cancer detection can be achieved through two distinct methods: positive screening or symptomatic identification. It is assumed that the implementation of a nationwide organized screening program, has played a significant role in improving patient survival rates for the last fifteen years. The main measure used to assess the effectiveness of breast cancer screening is the disease-specific survival rate after diagnosis, calculated using the Kaplan-Meier method. In order to estimate the survival improvement of patients detected following a screening, it is necessary to compare their survival time from diagnosis with those of patients detected following symptoms. However, screening introduces bias leading to an overestimation of survival of patients detected through screening.

## 1.2 Lead time bias

The lead-time bias is a well-known issue in breast cancer screening. Survival times from the diagnosis of patients following screening versus those following symptoms are not comparable. The reason is that the screening advanced the diagnosis in time. The lead time, that cannot be observed, is the duration between the date of diagnosis by screening and the hypothetical date of diagnosis by symptoms for screen-detected patients. The screening introduces an artificial extension of the survival for these patients, which is not observed in patients diagnosed by symptoms (see figure 1). This bias on the survival of patients diagnosed by screening has consequences. The longer survival observed in patients diagnosed by screening is not entirely due to the early detection, but also to the lead-time added by the screening process. Therefore, the actual benefit of screening in terms of improving patient survival is overestimated.

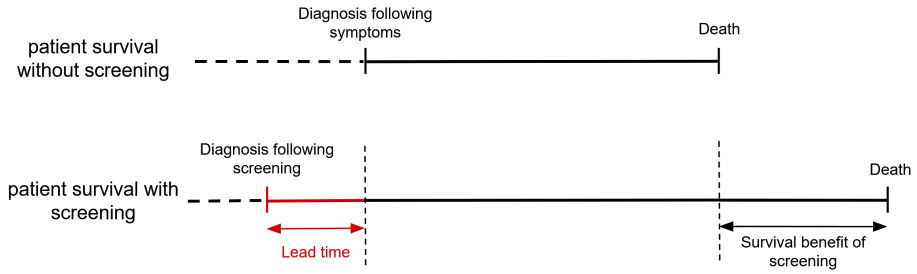



Figure 1: Illustration of the survival time comparison since diagnosis for a patient with or without screening.

# 2 Objective

The objective is to use two lead time correction approaches on real data and compare them. Ultimately, this project will result in the development of an R package, facilitating the application of these methods by epidemiologists.

# 3 Method

Two frequently used methods for correcting lead time have been compared: the first one is based on estimates from a multi-state Markov model, and the second one is based on a continuous tumor growth model. These two methods are applied in three steps. First, a model is estimated for the overall study population over the screening period, from their first mammography to their diagnosis. Then, the parameters of these models are used to determine a distribution of the additional time added due to the lead time bias. A time is then subtracted from the survival since diagnosis of each screen-detected patient.

## 3.1 Multi-state approach

### 3.1.1 Homogeneous Markov model

Duffy et al. (1995) proposed to model the tumor's progression with a three-states Markov model (white boxes in figure 2), each state corresponding to a phase in the natural history of the disease: no detectable tumor (state 1), the preclinical phase (state 2) when a tumor is detectable by screening but not symptomatic, the clinical phase (state 3) when the disease becomes symptomatic.

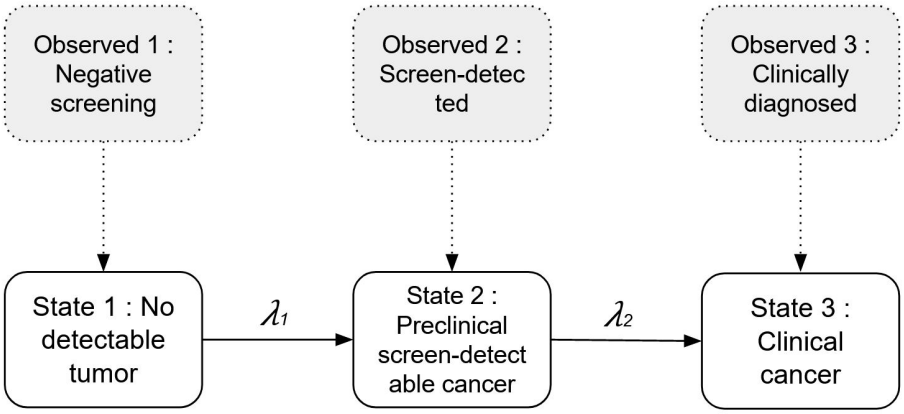



Figure 2: Three-state Markov model of the breast cancer natural history: the gray boxes represent the observed states, and the white boxes their underlying state in the model.

Negative mammography corresponds to an observation in state 1, where there is either no tumor or the tumor is not detectable. A positive screening mammography corresponds to an observation in state 2. The patient has no symptoms but the tumor is detectable. Diagnosis based on symptoms corresponds to an observation in state 3. The transition from state 1 to state 2 is interval-censored, meaning that the observations are in discrete time and do not provide information on the exact time of transition from one state to another. It is assumed that observations in state 3 occur at the exact time of transition into this state, which corresponds to diagnosis at the onset of symptoms and the clinical cancer phase in the natural history of the disease. Based on the definitions of the different states, no patient can be successively observed in states 2 and 3.

### 3.1.2 Breast density adjustment:

One of the main criticisms of the initial model (Duffy et al., 1998) is that it doesn't take into account patients' differences in cancer progression speed and therefore adjustments on sojourn time estimation. The underlying assumption with our adjustments is that model parameters and lead time correction are different depending on the four categories of breast density using the BI-RADS classification (Boyd et al., 2007; Choi et al., 2022). Transition

rates are estimated using the following formula with three parameters added per transition for the breast density indicator variables:

$$\lambda_j = exp(\beta_{0_j} + \beta_{1_j} br2 + \beta_{2_j} br3 + \beta_{3_j} br4)$$

for $j = 1, 2$, the transition.

### 3.1.3 Lead time bias correction

The correction for lead time based on the estimates of this model was published in 2008 by Duffy et al. The principle of this correction is to subtract from patients' survival time, the part of the sojourn time in preclinical phase in their post-diagnosis follow-up $s$. The sojourn time must follow an exponential distribution so the transition from preclinical state to clinical state has to be homogeneous in time. In the correction of Duffy et al., we define $t$, the time elapsed since the screening diagnosis, then there are two possibilities:
The patient passed away at the time of the latest news after diagnosis $t$:

$$E(s) = P(s \leq t)E(s|s \leq t) = \frac{1 - e^{-\lambda_2 t} - \lambda_2 t e^{-\lambda_2 t}}{\lambda_2(1 - e^{-\lambda_2 t})} \tag{1}$$

The patient is alive at the time of the latest news after diagnosis $t$:

$$E(s) = P(s \leq t)E(s|s \leq t) + P(s > t)t = \frac{1 - e^{-\lambda_2 t}}{\lambda_2} \tag{2}$$

Finally, the correction is applied to the survival of each screen-detected patient $i$ by subtracting the average amount of sojourn time in post-diagnosis follow-up $E(s)$.

$$T_{corrected,i} = T_{observed,i} - E(s)_i \tag{3}$$

## 3.2 Tumor growth approach

The second correction method, which is more recent and complex, involves estimating a continuous tumor growth model. The continuous tumor growth model allows us to estimate the volumetric growth of tumors over time until a detection date through symptoms, with the assumption of exponential growth. It will then be possible to estimate the necessary time for growth from the size at diagnosis after the screening to the hypothetical detection with symptoms.

### 3.2.1 Continuous tumor growth model

These models are based on a tumor growth function and are estimated from latent sub-models (Isheden et al., 2017; Strandberg et al., 2019):

$$V(x) = v_0 e^{\frac{x}{r}}$$

where $v_0$ is the tumor volume at onset. We assume tumors are spherical, and since we define onset as the point where the tumor diameter is 0.5mm, then $v_0 \approx 0.06mm^3$. This function allows expressing the growth rate $r > 0$ as a function of the volume at detection $v$, the age at detection $x$ that are observed.

In the model the individual likelihood contributions are based first on the probability of detecting a tumor of volume $v$. We use an exponential density function of the tumor volume at symptomatic detection ($V_{sym}|R = r \sim Exp(\eta r)$) marginalized on the gamma density function of the tumor growth rate ($R \sim Gamma(\tau_1, \tau_2)$).

The individuals contributions also take into account the probability of detecting the tumor at the screening detection and the probability of not detecting the tumor at the previous screening knowing the inverse growth rate $r$. For these probabilities the screening sensitivity is dependent on the tumor diameter and the BI-RADS category, and is assumed to be of logistic form.

$$S(d) = \frac{exp(\beta_0 + \beta_1 d + \beta_2 br2 + \beta_3 br3 + \beta_4 br4)}{1 + exp(\beta_0 + \beta_1 d + \beta_2 br2 + \beta_3 br3 + \beta_4 br4)}, \qquad d \geq 0$$

where d is the diameter of the tumor, which is observed at diagnosis, and which is calculated during previous screening using the tumor growth function.

### 3.2.2 Lead time bias correction

The model and its parameters will then be used to estimate a density function of the lead time conditionally on the tumor volume, the growth rate, and the screening sensitivity (Abrahamsson et al., 2020). The general idea is to create a customized density function for each patient tracking her follow-up from diagnosis to the date of the latest news and evaluating at each time point the probability of symptomatic detection. We denote L, a random variable representing a quantity of lead time, and $l$ a realization of this variable. It is assumed that $L = 0$ for patients detected symptomatically, and $L > 0$ for screen-detected. The lead time is the time between the screen-detection time point $t_{scr}$ and the hypothetical symptomatic detection time point $T_{sym}$ ($L = T_{sym} - t_{scr}$), expressed in the model using the tumor growth function. Mathematically, we express the lead time as follows:

$$V_{sym} = v_{scr} e^{(T_{sym} - t_{scr})/R} \iff L = Rlog(\frac{V_{sym}}{v_{scr}})$$

The density function of lead time is in the following form:

$$f_{L|V=v_{scr},H}(l) \propto \int_0^\infty f_{R|V=v_{scr}}(r) \cdot f_{L|V=v_{scr},R=r}(l) \cdot P(H|V = v_{scr}, R = r)dr \qquad (4)$$

- The first part of this integral is for the probability of having an inverse growth rate $r$ given the tumor volume at detection $V = v_{scr}$, calculated by using the conditional density function of the inverse growth rate.

- The second part is the expression of lead time in the model, given the tumor volume at detection $v_{scr}$ and the inverse growth rate $r$. We derive the cumulative density function of the volume at symptomatic detection with the previously mentioned relation $V_{sym} = v_{scr}e^{L/R}$.

- The third part, depending on the screening sensitivity, is the probability of negative screening results in the patient's mammography history $H$. For each mammography, the tumor diameter is backward calculated as before by using the inverse growth rate $r$.

Similarly to the correction made by Duffy et al. (2008), the time subtracted from survival should not exceed the observed survival time for a dead patient. The lead time distribution is then in a conditional truncated form (Abrahamsson et al., 2020). The final value $\hat{l}_i$ is estimated for each woman by taking the expectancy of her lead time density function. The correction on survival is performed for each screen-detected patient by subtracting the lead time estimate from the observed survival.

$$T_{corrected,i} = T_{observed,i} - \hat{l}_i \tag{5}$$

# 4 Results

The data used for the analysis come from the Gironde general cancer registry. Some patients were excluded from the analyses due to unknown detection method, missing data on tumor size or the breast density at diagnosis, or the presence of an in situ tumor, which did not verify the growth hypothesis of the model.

Figure 3 shows a comparison of patient survival by diagnostic method. The probability of survival at 5 years for a woman diagnosed by symptom is 0.85. After a diagnosis by screening, the probability of survival at 5 years is 0.95. The corrected survival curves are above the curve for symptomatic patients, reflecting the fact that screening is effective on survival after correction of the lead time. The correction based on a multi-state model is the strongest of the two, and the survival curve tends to approach that of symptomatic patients. The correction based on tumor growth remains fairly close to that of survival without correction, the decreases in survival probabilities remain of the same order of magnitude.

# 5 Discussion

The correction based on a multi-state model is stronger, possibly even too strong, compared to the one based on tumor growth, which is more comprehensive in estimating lead time. Applying either of these corrections during survival analysis is of crucial importance for estimating the effectiveness of screening. It is necessary to keep in mind that both methods have their own assumptions, strengths, and limitations.

The definition of states is advantageous in order to estimate sojourn times and incidence rates. However, the main issue with the three-state model is that it only allows for estimating sojourn time and not lead time.

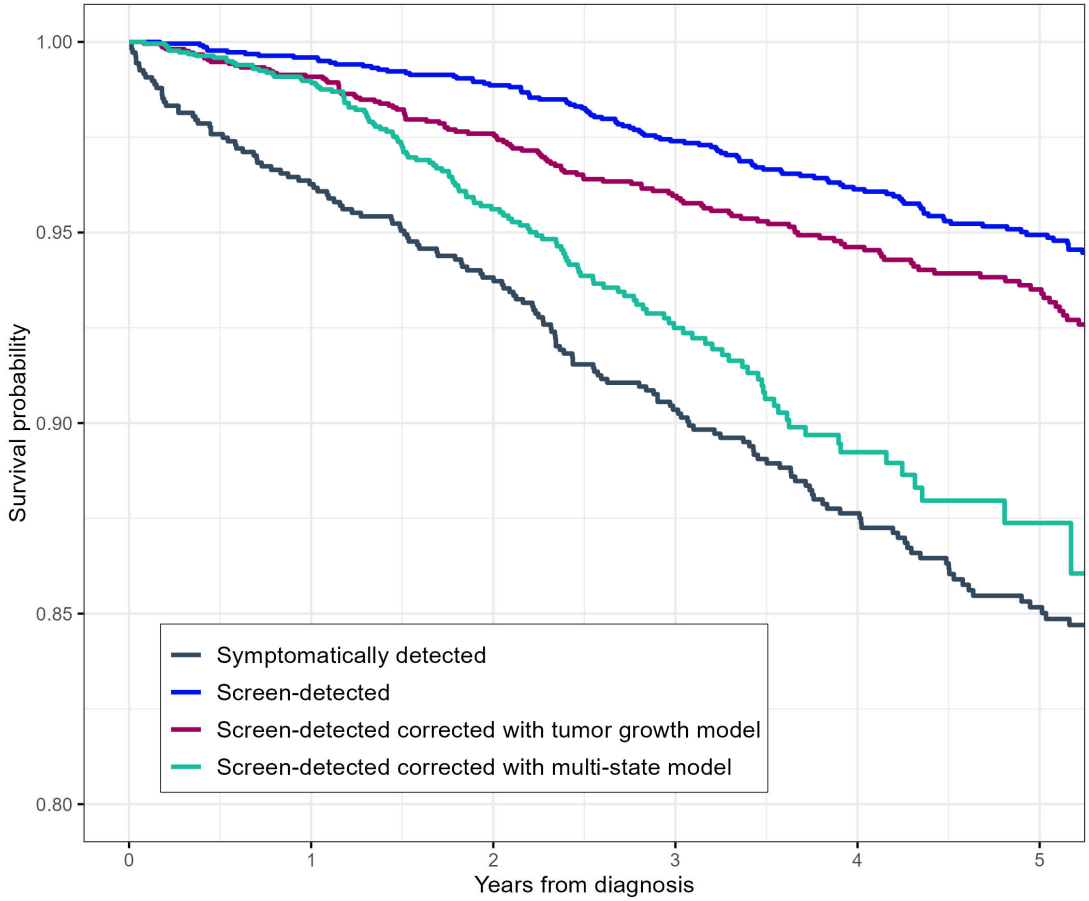



Figure 3: Survival comparison after breast cancer diagnosis for women detected by symptoms and screening according to different estimated lead time bias corrections. Symptomatic diagnosis: $n_{sym} = 1077$, screen-detected: $n_{scr} = 2212$.

The continuous tumor growth model and its associated correction directly address the limitations of the multi-state approach due to their assumptions. The distribution of lead time is personalized for each woman and adjusted on tumor size at diagnosis and screening sensitivity. However, some hypotheses constitute limitations. Particularly, the assumptions that tumors are spherical in shape and grow exponentially are not universally valid. Other shapes, such as ellipses, are possible, and tumor protrusions are common.

We assumed no modification in women's breast density over time. We were able to verify this through follow-ups of up to 10 years between the initial screenings in 2005 and the later diagnosis in 2015. However, the screening program is designed for 25 years, which is more than double the follow-up period. Therefore, breast density tends to decrease with age and, our assumption should be invalid.

# 6 Perspectives

Screening has a higher probability of detection for slow-growing tumors, which constitutes another bias. Extending a method towards selection biases linked to screening constitutes the continuation of this work.

# References


Abrahamsson, L., Isheden, G., Czene, K., Humphreys, K. (2020). Continuous tumour growth models, lead time estimation and length bias in breast cancer screening studies. *Statistical Methods in Medical Research*, 29(2), 374-395.

Boyd, N. F., Guo, H., Martin, L. J., Sun, L., Stone, J., Fishell, E., ... Yaffe, M. J. (2007). Mammographic density and the risk and detection of breast cancer. *New England journal of medicine*, 356(3), 227-236.

Choi, E., Suh, M., Jung, S. Y., Jung, K. W., Park, S., Jun, J. K., Choi, K. S. (2022). Estimating Age-Specific Mean Sojourn Time of Breast Cancer and Sensitivity of Mammographic Screening by Breast Density among Korean Women. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, 55(1), 136-144.

Duffy, S. W., Chen, H. H., Tabar, L., Day, N. E. (1995). Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in medicine*, 14(14), 1531-1543.

Duffy, S. W., Nagtegaal, I. D., Wallis, M., Cafferty, F. H., Houssami, N., Warwick, J., ... Lawrence, G. (2008). Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. *American journal of epidemiology*, 168(1), 98-104.

Isheden, G., Humphreys, K. (2019). Modelling breast cancer tumour growth for a stable disease population. *Statistical Methods in Medical Research*, 28(3), 681-702.

Strandberg, J. R., Humphreys, K. (2019). Statistical models of tumour onset and growth for modern breast cancer screening cohorts. *Mathematical Biosciences*, 318, 108270.

# Statistique robuste et détection d'anomalie

# Robust estimation in linear mixed effects models

Valérie Garès[1], Rik Lopuhaä[2] & Anne Ruiz Gazen[3]

[1] *Univ. Rennes INSA, CNRS, IRMAR - UMR 6625, Rennes, France,*
*valerie.gares@insa-rennes.fr*
[2] *Delft University of Technology*
[3] *Toulouse School of Economics*

**Résumé.** Les modèles linéaires à effets mixtes sont largement utilisés pour étudier des réponses corrélées notamment pour l'analyse de données longitudinales, de données de croissance ou des mesures répétées. Les estimateurs classiques de ces modèles, tels que les estimateurs du maximum de vraisemblance, sont basés sur des hypothèses de normalité et sont sensibles aux valeurs atypiques. Il est donc important d'explorer des estimateurs robustes dans ce contexte. Nous nous concentrons sur les modèles linéaires à effets mixtes équilibrés et proposons un aperçu des méthodes d'estimation robustes qui ont été étudiées et revisitées ces dernières années, tels que les estimateurs S, MM et l'estimateur tau composites. Lors de la présentation, nous rappellerons brièvement leur définition et leurs propriétés théoriques, et les comparerons via une étude par simulations.

**Mots-clés.** Estimateurs composites, Estimateurs robustes, Modèles linéaires mixtes, MM estimateurs, S estimateurs, tau estimateurs.

**Abstract.** Linear mixed effects models are widely used to study correlated responses, particularly for the analysis of longitudinal data, growth data or repeated measures. Conventional estimators of these models, such as maximum likelihood estimators, are based on assumptions of normality and are sensitive to outliers. It is therefore important to explore robust estimators in this context. We focus on balanced linear mixed effects models and provide an overview of robust estimation methods that have been studied and revisited in recent years, such as the S, the MM and the composite tau estimators, and their composite counterparts. We will briefly recall their definition and their theoretical properties, and we will compare them via a simulation study.

**Keywords.** Composite estimators, Linear mixed effects models, MM-estimators, Robust estimators, S-estimators, tau-estimators.

## 1 Introduction

Linear models with structured covariance matrices are widely used and provide a versatile approach for analyzing correlated responses, such as longitudinal data, growth data or repeated measurements. In such models, each subject $i$, $i = 1, \ldots, n$, is observed at $k_i$ occasions, and the vector of responses $\mathbf{y}_i$ is assumed to arise from the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i,$$

where $\mathbf{X}_i$ is the design matrix for the $i$th subject and $\mathbf{u}_i$ is a vector whose covariance matrix can be used to model the correlation between the responses. We consider a structured covariance matrix, that is, the matrix $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ is a known function of unknown covariance parameters combined in a vector $\boldsymbol{\theta} \in \mathbb{R}^l$.

The balanced linear mixed effects model is a well-known example of a linear model with a structured covariance matrix. We consider independent observations $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$ such that

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{j=1}^{r} \mathbf{Z}_j\gamma_{ij} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n. \tag{1}$$

For each subject $i = 1, \ldots, n$, $\mathbf{y}_i \in \mathbb{R}^k$ is the response vector and $\mathbf{X}_i \in \mathbb{R}^{k\times q}$ the known design matrix for the fixed effects. The vector $\boldsymbol{\beta} \in \mathbb{R}^q$ is the unknown fixed effects parameter. The $\mathbf{Z}_j$'s are known $k \times g_j$ design matrices for the random effects $\gamma_{ij} \in \mathbb{R}^{g_j}$, which are assumed to be independent mean zero random vectors with covariance matrix $\sigma_j^2\mathbf{I}_{g_j}$, for $j = 1, \ldots, r$. The error terms $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, belong to $\mathbb{R}^k$, are independent mean zero random vectors with covariance matrix $\sigma_0^2\mathbf{I}_k$, and are independent from the $\gamma_{ij}$'s. This means that we concentrate on models for which

$$\mathbf{V}(\boldsymbol{\theta}) = \sum_{j=1}^{r} \sigma_j^2\mathbf{Z}_j\mathbf{Z}_j^T + \sigma_0^2\mathbf{I}_k \text{ and } \boldsymbol{\theta} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_r^2).$$

Common estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are maximum likelihood estimators, derived under Gaussian assumptions for the random effects and for the error terms (see, e.g., Hartley and Rao, 1967; Laird and Ware, 1982). It is well-known that these estimators rely heavily on the Gaussian assumptions and are highly sensitive to outlying observations (see, e.g., Welsh and Richardson, 1997). To overcome this problem, several robust estimation methods have been proposed and studied in the last thirty years (see, e.g., Agostinelli and Yohai, 2016; Mason et al., 2021; Lopuhaä, 2023, for some recent overviews and studies).

Among the existing robust approaches, some robust estimators are aimed at resisting high proportions of outliers in different contamination models, and we will focus on such estimators. In the robust statistical literature, there exist two contamination models: the Classical (Tukey-Huber) Contamination Model (CCM) and the Independent Contamination Model (ICM). While CCM (also called "case-wise" contamination) considers that the contamination is at the level of the subjects or cases, ICM (also called "cell-wise" contamination) considers the possibility to contaminate data sets at the level of the cells. Initially, the robust statistics literature focused on the CCM context and proposed robust estimators that were able to cope with a proportion of outliers close to 50% without breaking down. However, such estimators may not be robust in the ICM context, since even a small fraction of contaminated cells may lead to more than 50% of contaminated cases.

The aim of this presentation is to give a short overview of some highly robust estimators in the CCM and ICM contexts, and compare their behavior on a simulation study. In Section 2, we introduce briefly some robust estimators we want to compare. In Section 3, we present the simulation scenarios we will consider. Results will be detailed and discussed during the presentation.

## 2 Robust estimators for balanced linear mixed effects models

The first robust estimators for linear mixed effects models were based on weighted versions of the likelihood function (see, e.g., Welsh and Richardson, 1997).

Then, the well-known S-estimators were extended to linear mixed effects models in Copt and Victoria-Feser (2006) (see also Heritier et al., 2009). S-estimators are smooth versions of the minimum volume ellipsoid estimator proposed in Rousseeuw (1985). They are called high-breakdown point estimators because they can resist to a large proportion of case-wise outliers without breaking down. The theoretical properties of S-estimators and in particular their asymptotic distribution, were revisited very recently in Lopuhaä et al. (2023) for general linear models with structured covariance.

S-estimators may also serve as initial estimators for MM estimators of the fixed effects parameter $\boldsymbol{\beta}$ proposed by Copt and Heritier (2007), and revisited recently by Lopuhaä (2023). MM-estimators are expected to be more efficient than S-estimators while maintaining a high breakdown point in the CCM context.

S and MM-estimators are not able to cope with case-wise contamination rate larger than 50% that may arise in the ICM context. To overcome this problem, Agostinelli and Yohai (2016) proposed a composite approach for tau-estimators (Yohai and Zamar, 1988) inspired by Lindsay (1988). The proposed composite tau estimator is highly robust not only under CCM but also under ICM.

Our objective is to compare the different estimators for CCM and ICM through a Monte Carlo study detailed below.

## 3 Simulation setup

We will propose a Monte Carlo study of the behavior of the above estimators for samples generated using models inspired by the simulation setup of Mason et al. (2021). We consider uncontaminated data and several types of contaminated data (i) in the random effects, (ii) in the measurement error terms, and (iii) in the design matrix of the fixed effects, according to CCM and ICM. We go beyond Mason et al. (2021) by considering not only ICM but also CCM, by taking larger proportions of contamination and by looking at two types of contamination in the design matrix of the fixed effects. Let us detail the generated data and the setup.

**Uncontaminated data.** Let us consider a linear mixed effects model with $\mathbf{y}_i$ in dimension $k = 4$ such that:

$$\mathbf{y}_i = 250\,\mathbf{1} + 10\,\mathbf{x}_i + \gamma_{0i}\mathbf{1} + \gamma_{1i}\mathbf{x}_i + \epsilon_i = \mathbf{X}\boldsymbol{\beta} + \gamma_i\mathbf{Z} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{2}$$

with $\mathbf{1}$ the vector of ones of dimension 4, $\mathbf{x}_i = (0, 1, 2, 3)^T$, the fixed effects $\boldsymbol{\beta} = (250, 10)^T$,

the random effects

$$\gamma_i = \begin{pmatrix} \gamma_{0i} \\ \gamma_{1i} \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 790 & -8.5 \\ -8.5 & 40 \end{pmatrix} \right), \quad \text{and the error terms} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, 400\mathbf{I}).$$

For the contaminated data, we assume that a proportion $(1-\delta)$ of subjects for the CCM, and cells for the ICM, is generated following the model defined in equation (2), while the remaining proportion $\delta$ of subjects for the CCM, and cells for the ICM, are outlying.

**Contamination of the errors.** Let us consider a shift $m_\epsilon > 0$.

- For CCM, $\boldsymbol{\epsilon}_i \sim (1-\delta)\mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, 400\,\mathbf{I} \right) + \delta\mathcal{N}\left( \begin{pmatrix} m_\epsilon \\ 0 \\ 0 \\ 0 \end{pmatrix}, 400\,\mathbf{I} \right)$.
- For ICM, $\boldsymbol{\epsilon}_{ij} \sim (1-\delta)\mathcal{N}(0, 400) + \delta\mathcal{N}(m_\epsilon, 0.25)$,

**Contamination of random slope effects.** Let us consider a shift $m_\gamma > 0$.

$$\gamma_i \sim (1-\delta)\mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 790 & -8.5 \\ -8.5 & 40 \end{pmatrix} \right) + \delta\mathcal{N}\left( \begin{pmatrix} 0 \\ m_\gamma \end{pmatrix}, \begin{pmatrix} 7.9 & -0.085 \\ -0.085 & 0.4 \end{pmatrix} \right)$$

**Contamination of the design matrix of the fixed effects.** We consider two different contamination frameworks. The first one is such that data are generated according to model (2) and, for a given $\alpha > 1$:

- For CCM, a proportion $\delta$ of $\mathbf{x}_i$ is replaced by $\alpha\mathbf{x}_i$.
- For ICM, a proportion $\delta$ of $x_{ij}$ is replaced by $\alpha x_{ij}$.

For the second framework, the design matrix of the uncontaminated data is such that the $\mathbf{x}_i$'s are generated independently and follow a standard Gaussian distribution in 4 dimensions.

- For CCM,
$$\boldsymbol{x}_i \sim (1-\delta)\mathcal{N}\left( \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{I} \right) + \delta\mathcal{N}\left( \begin{pmatrix} m_x \\ \vdots \\ m_x \end{pmatrix}, \mathbf{I} \right).$$
- For ICM, $x_{ij} \sim (1-\delta)\mathcal{N}(0,1) + \delta\mathcal{N}(m_x, 1)$.

In order to investigate the impact of the different contamination parameters, we will choose $\delta \in \{0, 5, 10, 20, 30\}$, the shifts $m_\epsilon \in \{-40, -80, -160\}$, $m_\gamma \in \{-40, -80, -160\}$, $m_x \in \{0.5, 1, 5, 10\}$, and the parameter $\alpha \in \{2, 5, 10, 50, 100\}$. For these different parameters, we will draw parallel boxplots for the maximum likelihood estimate and for the robust S, MM and composite tau estimates. We will compare the results and give some recommendations.

# References


Agostinelli, C. and Yohai, V. J. (2016). Composite robust estimators for linear mixed models. *Journal of the American Statistical Association*, 111(516):1764–1774.

Copt, S. and Heritier, S. (2007). Robust alternatives to the f-test in mixed linear models based on MM-estimates. *Biometrics*, 63(4):1045–1052.

Copt, S. and Victoria-Feser, M.-P. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, 101(473):292–300.

Hartley, H. O. and Rao, J. N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2):93–108.

Heritier, S., Cantoni, E.and Copt, S., and Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics*. John Wiley & Sons.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lindsay, B. G. (1988). Composite likelihood methods. *Comtemporary Mathematics*, 80(1):221–239.

Lopuhaä, H. P. (2023). Highly efficient estimators with high breakdown point for linear models with structured covariance matrices. *Econometrics and Statistics*.

Lopuhaä, H. P., Gares, V., and Ruiz-Gazen, A. (2023). S-estimation in linear models with structured covariance matrices. *Annals of Statistics*, 51(6):2415–2439.

Mason, F., Cantoni, E., and Ghisletta, P. (2021). Parametric and semi-parametric bootstrap-based confidence intervals for robust linear mixed models. *Methodology*, 17(4):271–295.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(283-297):37.

Welsh, A. and Richardson, A. (1997). 13 approaches to the robust estimation of mixed models. *Handbook of statistics*, 15:343–384.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical association*, 83(402):406–413.

# Robustesse de la profondeur scatter

Gaëtan Louvet [1] & Germain Van Bever [2]

[1] *Université libre de Bruxelles, Belgique, gaetan.louvet@ulb.be*
[2] *Université libre de Bruxelles, Belgique, germain.van.bever@ulb.be*

**Résumé.** La profondeur statistique fournit des outils non paramétriques robustes pour analyser les distributions. En effet, les fonctions de profondeur mesurent l'adéquation entre les paramètres d'une distribution et les mesures de probabilité sous-jacentes. Dans le cas *position*, un exemple de telle mesure est la profondeur de demi-espace de Tukey, dont les propriétés de robustesse ont déjà été largement étudiées. Récemment, des notions de profondeur pour les paramètres de *dispersion* ont été définies et étudiées. Les propriétés de robustesse de ces fonctions de profondeur scatter restent toutefois largement inconnues. Dans cet exposé, nous présentons différents résultats concernant la robustesse de la profondeur scatter ainsi que de la médiane associée. Nous en dérivons la fonction d'influence et le point de rupture, et déduisons la distribution asymptotique des versions empiriques de celles-ci.

**Mots-clés.** Profondeur scatter, robustesse, fonction d'influence, point de rupture, distribution asymptotique.

**Abstract.** Statistical depth provides robust nonparametric tools to analyze distributions. Depth functions indeed measure the adequacy of distributional parameters to underlying probability measures. In the *location* case, the celebrated halfspace depth has been widely studied, and its robustness properties are amply discussed. Recently, depth notions for *scatter* parameters have been defined and studied. The robustness properties of this latter depth function remain, however, largely unknown. In this talk, we present several results regarding the scatter depth function and its associated scatter median, including the influence function, the breakdown point and the asymptotic distribution.

**Keywords.** Scatter depth, robustness, influence function, breakdown point, asymptotic distribution.

## 1 Profondeur scatter

Génériquement, une fonction de profondeur $D(\cdot, \mathrm{P}) : \theta \mapsto \mathrm{D}(\theta, \mathrm{P})$ associe à tout paramètre $\theta$ une mesure de l'adéquation de celui-ci pour la distribution de probabilité P sous-jacente. Cet outil nonparamétrique permet typiquement de construire des méthodologies robustes pour l'estimation ou le test dans des distributions paramétriques.

Dans le cas *position*, la profondeur de demi-espace de Tukey (1975) est définie comme

$$HD_{\mathrm{loc}}(\theta, \mathrm{P}) = \inf_{||u||=1} \mathrm{P}\left[u'(X - \theta) \geq 0\right].$$

Ses propriétés de robustesse ont déjà été largement étudiées: fonction d'influence (Romanazzi (2001)), point de rupture (Liu *et al.* (2017)) et comportement asymptotique (Massé (2002)).

Récemment, des notions de profondeur pour les paramètres de *dispersion* ont été définies et étudiées. Les propriétés de robustesse de ces fonctions de profondeur *scatter* restent toutefois largement inconnues. Dans ce document, nous examinons principalement la robustesse de la profondeur du demi-espace de scatter introduite par Chen *et al.* (2018), dont l'expression est donnée par

$$HD_{\text{sc}}(\Sigma, \text{P}) = \inf_{||u||=1} \min\left(\text{P}\left[|u'(X - T_{\text{P}})| \leq \sqrt{u'\Sigma u}\right], \text{P}\left[|u'(X - T_{\text{P}})| \geq \sqrt{u'\Sigma u}\right]\right),$$

et où $T_{\text{P}}$ est un estimateur de position. Nous considérons également la médiane de scatter associée, définie, pour $\mathcal{M}^p$ l'ensemble des matrices symétriques définies positives, comme

$$S_{HD}(\text{P}) = \text{argmax}_{\Sigma \in \mathcal{M}^p} HD_{\text{sc}}(\Sigma, \text{P}).$$

# 2 Fonction d'influence

La fonction d'influence mesure la sensibilité d'une fonctionnelle $S$ à l'ajout d'une faible contamination sur la distribution P. Elle se définit formellement comme

$$\text{IF}(z; S(\text{P})) = \lim_{\varepsilon \to 0^+} \frac{S(\text{P}_{\varepsilon,z}) - S(\text{P})}{\varepsilon} = \frac{\partial}{\partial \varepsilon} S(\text{P}_{\varepsilon,z})\Big|_{\varepsilon=0^+},$$

où $\text{P}_{\varepsilon,z} = (1-\varepsilon)\text{P} + \varepsilon\Delta(z)$ correspond à la distribution contaminée et $\Delta(z)$ est la distribution de Dirac en $z \in \mathbb{R}^p$. Les résultats obtenus concernant la fonction d'influence de la profondeur scatter se distinguent selon l'estimateur de position utilisé et/ou les hypothèses de continuité sur la distribution sous-jacente. Nous décrivons ceux-ci de manière informelle ci-dessous.

- Dans le cas où le paramètre de position est supposé connu, la fonction d'influence de la profondeur scatter est bornée et s'écrit comme

  $$\text{IF}(z; HD_{\text{sc}}(\Sigma, \text{P})) = I[z \in A] - HD_{\text{sc}}(\Sigma, \text{P}),$$

  pour $A = A(\text{P})$ un ensemble dépendant de la distribution P.

- Pour des distributions absolument continues et un estimateur de position $T_{\text{P}}$ quelconque, nous trouvons, sous certaines hypothèses sur P et $T_{\text{P}}$, pour une certaine fonction $g$, bornée si $\text{IF}(.; T_P)$ est bornée, que

  $$-HD_{\text{sc}}(\Sigma, \text{P}) + g(\text{IF}(z; T_{\text{P}})) \leq \text{IF}(z; HD_{\text{sc}}(\Sigma, \text{P})) \leq 1 - HD_{\text{sc}}(\Sigma, \text{P}) + g(\text{IF}(z; T_{\text{P}})).$$

- Dans le cas de distributions discrètes, il est impossible de dégager une expression générale explicite pour la fonction d'influence. Dans le cas où nous choisissons la médiane de demi-espace et que cette dernière est unique, nous retrouvons

  $$\text{IF}(z; HD_{\text{sc}}(\Sigma, \text{P})) = I[z \in A] - HD_{\text{sc}}(\Sigma, \text{P}).$$

L'obtention de la fonction d'influence de la médiane scatter en toute généralité est également impossible. Cependant, dans le cas de distributions elliptiques de densité radiale $f$ et de matrice de dispersion $\Sigma_*$, nous trouvons, pour certaines valeurs de $z$,

$$\mathrm{IF}(z; S_{HD}(\mathrm{P})) = -\Sigma_*/(2f(1)).$$

Pour des distributions plus générales, il est également possible d'obtenir des bornes sur la fonction d'influence.

# 3 Point de rupture

Le point de rupture d'une fonctionnelle est informellement défini comme le nombre d'observations à modifier avant que celle-ci n'atteigne sa ou ses bornes. Dans le cas de la profondeur, nous considérons donc, pour une jeu de données $X$, les points de rupture *par substitution*

$$\mathrm{BP}(HD_{\mathrm{sc}}(\Sigma, X)) = \min_{1\leq m\leq n} \{\frac{m}{n} : \sup_{X^m} HD_{\mathrm{sc}}(\Sigma; X^m) = 0\},$$

et

$$\mathrm{BP}(S_{HD}(X)) = \min(\mathrm{BP}^0(S_{HD}(X)), \mathrm{BP}^\infty(S_{HD}(X))),$$

où

$$\mathrm{BP}^0(S_{HD}(X)) = \min_{1\leq m\leq n} \{\frac{m}{n} : \inf_{X^m} \lambda_p(S_{HD}(X^m)) = 0\},$$

et

$$\mathrm{BP}^\infty(S_{HD}(X)) = \min_{1\leq m\leq n} \{\frac{m}{n} : \sup_{X^m} \lambda_1(S_{HD}(X^m)) = \infty\}$$

et où $X^m$ représente $X$ où $m$ données ont été arbitrairement modifiées et où $\lambda_1(S)$ et $\lambda_p(S)$ désignent, respectivement, la plus grande et la plus petite valeur propre de la matrice $S$. Les résultats sont les suivants:

- Lorsque l'estimateur de position $T_P$ est connu et fixé, la définition de profondeur implique que $\mathrm{BP}(HD_{\mathrm{sc}}(\Sigma, X)) = HD_{\mathrm{sc}}(\Sigma, X)$. Le cas position non fixé dépendra de la fonctionnelle de position utilisée.
- En notant $M = n \max_\Sigma HD_{\mathrm{sc}}(\Sigma, X)$, nous trouvons comme borne inférieure

  $$\mathrm{BP}^0(S_{HD}, X) \geq (\frac{\lceil M/2\rceil - p}{n})^+ \qquad \text{et} \qquad \mathrm{BP}^\infty(S_{HD}, X) \geq \frac{\lceil M/2\rceil}{n}.$$

  Asymptotiquement, le point de rupture atteint donc, dans le pire des cas, une valeur de $\max_\Sigma HD(\Sigma, \mathrm{P})/2$ ($= 0.25$ pour des distributions elliptiques par exemple).

# 4 Distribution asymptotique

Enfin, nous dérivons la distribution asymptotique de la médiane scatter pour des distributions continues P. Notons $\Sigma_0$ la médiane scatter sous P. Soit $G$ un vecteur aléatoire de même distribution que la distribution asymptotique de $\sqrt{n}(T_n - T_0)$. Soit $x : S^{p-1} \to \mathbb{R}$ un processus stochastique défini sur la sphère unité par, pour tout $u \in S^{p-1}$, $x(u) = K_1(u)u'G + \nu_F A_i(u, \Sigma_0, T_0)$, où $A_i$ un sous espace et $\nu_F$ un pont brownien. Sous certaines hypothèses sur P et pour certains sous-ensemble $V^i(0)$ et $V^o(0)$ de $S^{p-1}$, nous trouvons

$$\sqrt{n}(\Sigma_n - \Sigma_0) \to \arg\max_{\Sigma} \min(\inf_{V^i(0)} K_2(u)u'\Sigma u + x(u), \inf_{V^o(0)} -(K_2(u)u'\Sigma u + x(u))),$$

à condition que le maximum de la fonction ci-dessus est unique.

# Bibliographie


Tukey, J.W. (1975) Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, pp. 523-531

Romanazzi, M. (2001) Influence function of halfspace depth. *J. Multivariate Anal.* 77(1), pp. 138-161

Massé, J.-C. (2002) Asymptotics for the Tukey Median. *J. Multivariate Anal.* 81(1), pp. 286-300

Liu, X., Zuo, Y. et Wang, Q. (2017) Finite sample breakdown point of Tukey's halfspace median. *Sci. China Math.* 60, pp. 861-874

Chen, M., Gao, C. et Ren, Z. (2018) Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* 46(5), pp. 1932-1960

# Election Robustness Index

Jean-Baptiste Aubin[1], Léo Dort[1], and Antoine Rolland[2]

[1]Univ Lyon, INSA Lyon, UJM, UCBL, ECL, ICJ, UMR5208,
69621 Villeurbanne, France
[2]ERIC EA 3083, Université de Lyon, Université Lumière Lyon 2,
5 Pierre Mendès France, 69596 Bron Cedex, France

February 19, 2024

**Abstract**

Voting rules like majority judgement, range voting or approval voting are based on evaluations. In this context, we propose a new index giving the "Robustness of the election" of the winner of the voting process. This index is based on properties of depth contours.
Keywords : Depth functions, Depth contours, Evaluation based voting rules, Election Robustness Index.

## 1 Introduction

Representative democracy is widely present in developed countries. It heavily relies on elections during which representatives of the people (mayors, senators, deputies, president, etc.) are elected. The question of the legitimacy of elected officials is regularly addressed. In this work, we propose to provide a tool for assessing this legitimacy. This Election Robustness Index is developed within the framework of elections based on evaluations, specifically in the context of deepest voting linked to weighted $L^p$ depth functions (see Zuo [7]). These voting processes encompass the methods of majority judgment, range voting, and approval voting, the three most common voting techniques based on evaluations.

Initially, we recall what deepest voting entails, then we introduce the Election Robustness Index for the winning candidate, as well as associated indices for unsuccessful candidates. Lastly, we examine a set of properties verified by the index and we study a case using simulated data.

## 2 Weighted $L^p$ deepest voting

Deepest Voting (see Aubin et al. [1]) is a family of social decision functions based on evaluations. Let consider in the following that we have $n$ voters and $d$ candidates, and each

voter give a grade to each candidate. Without loss of generality, we suppose that these grades are in $[0;1]$. Each voter can then be seen as a point in $[0;1]^d$ whose components are the grades for each candidate. The set of all the voters' grades is then a point cloud. The key idea of Deepest Voting is to consider the grades of the *most central* voter of the cloud. This innermost (possibly imaginary) voter can be seen as the most representative of all the voters of the cloud, so that his preferences should meet the largest possible consensus among the other voters. The associated social decision function simply gives the grades of this innermost voter as output.

Quoting Liu et al. [5]: "associated with a given distribution $F$ on $\mathbb{R}^d$, a depth function is designed to provide a $F$-based center-outward ordering (and thus a ranking) of points $x$ in $\mathbb{R}^d$ . High depth corresponds to *centrality*, low depth to *outlyingness*". In other words, a depth function takes high (positive) values at the middle of a point cloud and vanishes out of it (see Zuo and Serfling [8] for a rigorous definition of a depth function). The intuitive key idea of Deepest Voting is enriched by a large choice of depth functions.

Let us introduce in particular the weighted $L^p$ depths. Consider a distribution of $n$ points $\Phi_n := \big(\Phi(\cdot,1),\dots,\Phi(\cdot,n)\big)$ in $\mathbb{R}^d$. The weigthed $L^p$ depth at a point $x \in \mathbb{R}^d$ is defined by Zuo [7] as:

$$wL^pD(x) = \frac{1}{1+\frac{1}{n}\sum_{j=1}^{n}\omega\big(\|\Phi_n(\cdot,j)-x\|_p\big)}$$

where $p>0$, $\omega$ is a non-decreasing and continuous function on $[0,\infty)$ with $\omega(\infty-)=\infty$ and $\|x-x'\|_p = \left(\sum_{i=1}^{d}|x_i-x'_i|^p\right)^{1/p}$. If $\omega : x \to x^p$, then for $x=(x_1,\dots,x_d)$,

$$wL^pD(x) = \frac{1}{1+\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{d}|\Phi_n(i,j)-x_i|^p}.$$

Not considering tie-breaking procedure, Majority Judgment and Range Voting (and its particular case, the approval voting) then correspond respectively to $wL^1$ Deepest Voting and $wL^2$ Deepest Voting. These are the three most famous voting processes based on evaluations.

The approval voting (see Brams and Fishburn [3] for a complete study) is maybe the most famous of these methods: each voter evaluates candidates on a scale of 2 gradings, which is the simplest possible scale. The voter gives a 1 if the candidate is acceptable, else a 0. The voter can then votes for several candidates (even all of them), or none of them accordingly to his convictions. Note that this method is very simple to apply in practice.

The two other methods are based on more nuanced classes of grading, which can be continuous or on a discrete scale. With the range voting proposed by Smith [6], the winner is the candidate with the highest average grade. With the majority judgment introduced by Balinski and Laraki [2], the winner is the candidate with the highest median grade. Note that the tie-break situation is taken into account for example in Fabre [4].

We visualise on figure 1 the example of the $wL^2$ depth function given in table 1.

| voter | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $v_8$ | $v_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $c_1$ | 0.30 | 0.10 | 0.70 | 0.70 | 0.65 | 0.70 | 0.20 | 0.15 | 0.20 |
| $c_2$ | 0.20 | 0.90 | 0.10 | 0.16 | 0.10 | 0.14 | 0.36 | 0.30 | 0.34 |

Table 1: Example of a distribution $\Phi_9$ of grades given by 9 voters on 2 candidates.

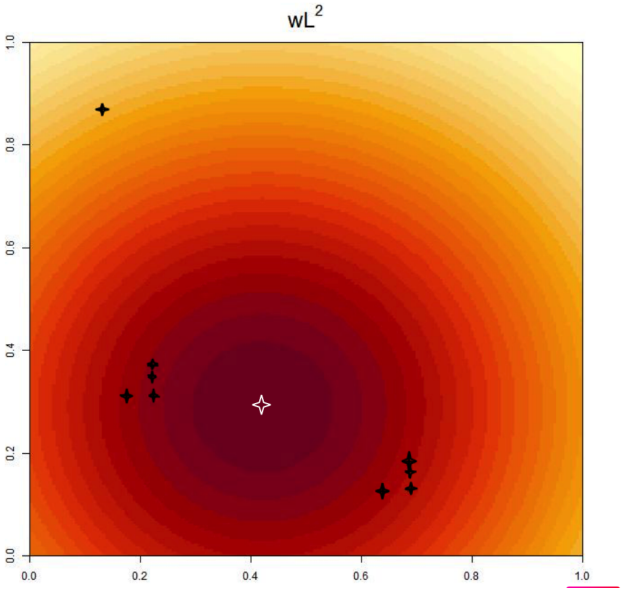



Figure 1: Example of $wL^2$ depth function on a data set. Horizontal axes give the grade for candidate $c_1$ and vertical axes for candidate $c_2$. Each black cross corresponds to a voter. The white cross corresponds to the deepest point.

# 3 Election Robustness Index for weighted $L^p$ deepest voting

The notion of $\alpha-$trimmed region, proposed in Zuo and Serfling [8], is introduced to define an Election Robustness Index.

**Definition 1.** *For a given depth function $D$ and a discrete distribution $\Phi_n$ and for $\alpha > 0$, we call*

$$D^\alpha(\Phi_n) := \left\{x \in [0;1]^d : D(x;\Phi_n) \geq \alpha\right\}$$

*the corresponding $\alpha-$trimmed region.*

Zuo and Serfling say that "The $\alpha-$trimmed region of a depth $D$ (exhibits the structure of underlying multivariate distribution and) reveals the shape of multivariate datasets".

Note that for $\alpha_1 \geq \alpha_2$, $D^{\alpha_1}(\Phi_n) \subseteq D^{\alpha_2}(\Phi_n)$. It means that the $\alpha-$trimmed regions are nested. On figure 1, the $\alpha-$trimmed regions are the nested areas of same color. The deepest

point for the w$L^2$ depth has coordinates (0.41, 0.29), inside the smallest $\alpha-$trimmed region, corresponding to the means componentwise.

Let denote by $D_p^\alpha$ the $\alpha-$trimmed regions for a $wL^p$ depth. Zuo and Serfling [8] show that $D_p^\alpha$ are compact for a distribution $\Phi$ continuous.

Let introduce the notion of preferential area $\mathcal{A}_i$ for the candidate $c_i$ as follows :

$$\mathcal{A}_i := \big\{x = (x_1, \ldots, x_d) \in [0;1]^d : \forall j \neq i,\ x_i \geq x_j\big\}.$$

Any set of preferences $x$ included in $\mathcal{A}_i$ has in common that the prefered candidate is $c_i$. Moreover, let consider the frontier of $\mathcal{A}_i$, denoted $\partial\mathcal{A}_i$, and defined as the set of points $x = (x_1, \ldots, x_d) \in [0;1]^d$ such that $x \in \mathcal{A}_i$ and it exists $j \neq i$ with $x_j = x_i$.

For example, in the case of 2 candidates, $\mathcal{A}_1$ is the set of $x = (x_1, x_2)$ such that $x_1 \geq x_2$. $\mathcal{A}_1$ is the area under its frontier $\partial\mathcal{A}_1 = \big\{x = (x_1, x_2) : x_1 = x_2\big\}$ materialized by the red line in Figure 2.

The Election Robustness Index represents the "centrality" of the deepest point inside the preferential area of the winner. It represents the "distance" of the deepest point from the preferential area of a loser.

**Definition 2.** *Given the preferences $\Phi_n$ and the deepest point associated to the $wL^p$ deepest voting $x_p^*$. Let denote $\partial\mathcal{A}_i$ the frontier of $\mathcal{A}_i$ and $\mathbb{1}_{\{x_p^* \in \mathcal{A}_i\}}$ the indicator function which gives 1 if $x_p^* \in \mathcal{A}_i$ and 0 elsewhere. Then the Election Robustness Index of candidate $c_i$ denoted $ERI_p(i)$ is given by :*

$$ERI_p(i) = \left(2\mathbb{1}_{\{x_p^* \in \mathcal{A}_i\}} - 1\right)\left(1 - \frac{\max\limits_{x \in \partial\mathcal{A}_i} D_p(x)}{D_p(x_p^*)}\right)$$

Note that only one candidate is associated to a positive $ERI_p$ (the winner of the w$L^p$ deepest voting procedure). All the others candidates have a negative $ERI_p$. The $ERI_p$ has always an absolute value lower than 1. The greater the $ERI_p$ is, the more robust the election is. In case of ex-aequos, their respective $ERI_p$ are null.

**Proposition 1.** *The previous properties are summarized in the following points :*

- *If $\exists!\ \ i_0$ such that $x_p^* \in \mathcal{A}_{i_0}$, then $ERI_p(i_0) > 0$ and $\forall i \neq i_0, ERI_p(i) < 0$ and $\max\limits_{i \neq i_0} ERI_p(i) = -ERI_p(i_0)$.*
- *If $\exists i_0 \neq i_1$ such that $x_p^* \in \mathcal{A}_{i_0} \cap \mathcal{A}_{i_1}$, then $ERI_p(i_0) = ERI_p(i_1) = 0$.*
- *If all elements of $\Phi_n$ are identical and included in the interior of $\mathcal{A}_{i_0}$, then $ERI_p(i_0) = 1$ and $\forall i \neq i_0, ERI_p(i) = -1$.*
- *For all $i$, $-1 \leq ERI_p(i) \leq 1$ and the winner is the only candidate with positive $ERI_p$.*

Next proposition is dedicated to a close form of the $ERI_2$ in the case of 2 candidates.

**Proposition 2.** *Let consider a matrix $\Phi_n$ of $n$ grades on 2 candidates and the $wL^2$ depth then, if $d := \left(\frac{\overline{\Phi_n(2,\cdot)}-\overline{\Phi_n(1,\cdot)}}{2}\right)^2$, then, if we assume that candidate 1 is the winner,*

$$ERI_2(1) = \frac{2d}{1 + var\big(\Phi_n(1,\cdot)\big) + var\big(\Phi_n(2,\cdot)\big) + 2d} = -ERI_2(2).$$

The application of the previous proposition to Table 1 leads to $d = 4.10^{-3}$ and $ERI_2(1) = 6.10^{-3}$. The Election Robustness Index is very small which illustrates the narrow victory of candidate 1. Moreover, Figure 2 visualises the largest deepest contour included in $\mathcal{A}_1$.

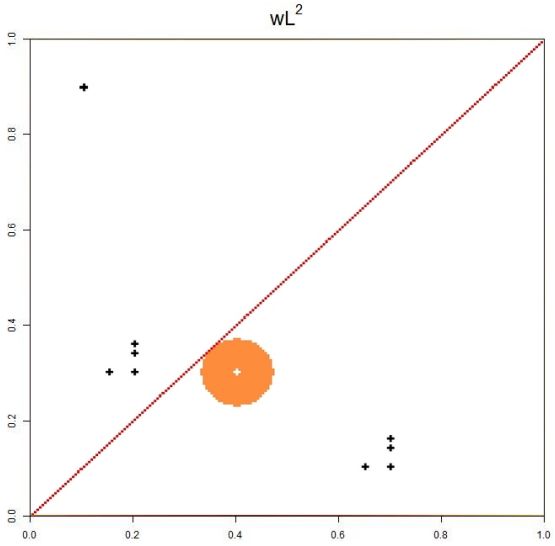



Figure 2: Example of the largest $\alpha-$trimmed region associated to a $wL^2$ depth function included in $\mathcal{A}_1$ in orange. Horizontal axes give the grade for candidate $c_1$ and vertical axes for candidate $c_2$. Each black cross corresponds to a voter, the white cross in the middle of the orange area is the deepest point.

Note to conclude that the choice of p in the w$L^p$ depth is crucial. For example, p=1 or a p large enough would lead to the election of candidate 2 in Table 1.

# References

[1] Jean-Baptiste Aubin, Irène Gannaz, Samuela Leoni, and Antoine Rolland. Deepest voting: A new way of electing. *Mathematical Social Sciences*, 116:1–16, 2022.

[2] Michel Balinski and Rida Laraki. A theory of measuring, electing and ranking. *Proceedings of the National Academy of Sciences*, 104(21):8720–8725, 2007.

[3] Steven J. Brams and Peter C. Fishburn. *Approval voting*. Springer, second edition, 2007.

[4] Adrien Fabre. Tie-breaking the highest median: alternatives to the majority judgment. *Social Choice and Welfare*, 56(1):101–124, 2021.

[5] Regina Y. Liu, Robert Serfling, and Diane L. Souvaine, editors. *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*.

[6] Warren D. Smith. Range voting, 2000.

[7] Yijun Zuo. Robustness of weighted Lp–depth and Lp–median. *Allgemeines Statistisches Archiv*, 88(2):215–234, 2004.

[8] Yiun. Zuo and Robert Serfling. General notions of statistical depth function. *Annals of Statistics*, 28:461–482, 2000.

# Détection d'anomalies dans des Données Mixtes : Évaluation des performances selon les types d'anomalies détectés

Houda GADACHA[1] & Patricia KUBICKI[2]

[1] *UTAC, France, houda.gadacha@utac.com*
[2] *UTAC, France, patricia.kubicki@utac.com*

**Résumé.**

Dans cette étude, notre objectif consiste à détecter des anomalies dans des données contenant à la fois des variables quantitatives et qualitatives. La plupart des méthodes de détection d'anomalies sont conçues uniquement pour les données quantitatives. Nous proposons d'utiliser une Analyse Factorielle sur Données Mixtes (AFDM) pour extraire des composantes principales numériques. Ces composantes sont ensuite utilisées pour la détection d'anomalies. Nous évaluons la performance de cette approche en utilisant des données simulées comportant différents types d'anomalies : globales, locales, rares et mixtes. Notre objectif est de déterminer les types d'anomalies détectés par chaque modèle.

**Mots-clés.** Détection d'anomalies, données mixtes, analyse factorielle des données mixtes, types d'anomalies.

**Abstract.**

In this paper, our aim is to detect anomalies in datasets containing both numerical and categorical attributes. Most outlier detection methods are designed only for numerical data. We propose using a Factor Analysis of Mixed Data (FAMD) to extract principal components. These components are then used for outlier detection. We evaluate the performance of this approach using simulated data containing various anomaly types : global, local, rare and mixed outliers. Our goal is to determine anomaly types detected by each model.

**Keywords.** Anomaly detection, mixed data, factor analysis on mixed data, outlier types.

## 1 Introduction

Selon Hawkins (1980), une observation est considérée comme une anomalie par rapport au reste de la population lorsqu'elle est générée par un mécanisme différent. En d'autres termes, une anomalie est une observation qui se distingue significativement des autres données.

La détection d'anomalies est cruciale dans divers domaines tels que la lutte contre la fraude à l'assurance, la détection des maladies et la sécurité informatique. Leur application permet de repérer des comportements suspects et d'améliorer la robustesse des modèles statistiques.

La détection d'anomalies suscite un intérêt marqué dans la littérature comme en témoignent les travaux de Togbe et al (2020), Liu et al (2008), Orair et al (2010), Syarif et al (2012), Breunig et al (2000), Münz et al (2007), et Knorr et al (1999). Bien que de nombreuses méthodes aient été proposées, la plupart se concentrent sur les données numériques. La détection d'anomalies pour les données mixtes, composées à la fois de variables quantitatives et qualitatives, a reçu moins d'attention. En pratique, certaines techniques de pré-traitement comme one hot encoding (OHE) sont souvent utilisées pour convertir les variables qualitatives en forme quantitative, permettant ainsi l'application de méthodes de détection d'anomalies adaptées aux données numériques.

Il existe principalement deux types d'anomalies en données quantitatives : globales et locales. Les anomalies globales sont des observations dont la valeur est significativement différente de celles de la majorité des observations. Ces anomalies se trouvent en dehors de la plage normale de l'ensemble de données. Les anomalies locales quant à elles ne sont pas nécessairement des valeurs extrêmes. Elles se situent dans la plage normale de l'ensemble de données mais sont considérées comme anormales par rapport à leur voisinage Alghushairy et al (2021). Ces deux types d'anomalies nécessitent des méthodes de détection différentes. Dans les données qualitatives, les anomalies sont appréhendées différemment. Certains articles sont focalisés dans la détection des catégories rares ou inattendues par rapport à la distribution normale des données Koufakou et al (2007) et Rokhman et al (2016). D'autres articles cherchent l'incompatibilité dans les combinaisons de modalités quand il s'agit de plusieurs variables qualitatives Taha et al (2016). Ainsi, certaines combinaisons de modalités peuvent être considérées comme des anomalies si elles sont inhabituelles ou très peu probables.

Dans cet article, nous poursuivons nos travaux portant sur notre approche novatrice pour la détection d'anomalies dans des jeux de données mixtes Gadacha et al 2023. La méthode repose sur l'application préalable d'une analyse factorielle de données mixtes AFDM suivi de l'analyse des facteurs numériques obtenus pour détecter les anomalies. Cette approche offre la possibilité d'utiliser tous les algorithmes conçus pour les données quantitatives. Nous comparons ces résultats à ceux d'une autre approche classique de pré-traitement de données qui consiste à créer une indicatrice binaire pour chaque modalité. Nous évaluons les performances des modèles et approches dans la détection des différents types d'anomalies.

Ce document est structuré de la manière suivante : dans la section 2 nous présenterons de manière concise l'état de l'art des modèles de détection d'anomalies, en mettant particulièrement l'accent sur ceux qui ont été utilisés dans nos expériences. La section suivante sera consacrée à l'explication de notre approche pour traiter les variables mixtes dans le but de détecter les anomalies. Ensuite, dans la section 4, nous analyserons l'efficacité des différents traitements en termes d'AUC et de types d'anomalies détectés.

# 2 Etat de l'art

D'après Togbe et al (2020), les méthodes de détection d'anomalies peuvent être regroupées en cinq grandes familles : La première concerne les techniques basées sur les tests statistiques pour identifier les observations qui s'écartent significativement de la distribution normale des

données (ex: Z-score, le test de Grubbs Aggarwal (2017)). La deuxième famille est composée de méthodes de clustering qui regroupent les observations en classes et sont utilisées pour la détection d'anomalies par l'identification des individus trop éloignés (ex: $K$-means, $K$-means ++ Arthur et al (2007), Velmurugan et al (2010), $K$-medoids, Density based spatial clustering of applications with noise DBSCAN Kriegel et al (1996)). La troisième famille concerne les méthodes des plus proches voisins qui identifient l'anomalie en comparant ses caractéristiques à celles de ses voisins les plus proches (ex: $K$-Nearest Neighbors KNN, Local Outlier Factor LOF Breunig et al (2000)). Dans la quatrième famille, nous trouvons les techniques basées sur le deep learning qui utilisent des réseaux de neurones profonds pour apprendre des représentations des données et détecter les anomalies à partir de ces représentations (ex: Auto-encoders AE et OneClass Neural Networks OCNN Chalapathy et al (2019)). La dernière famille comprend les autres méthodes qui ne rentrent pas directement dans les catégories précédentes comme Isolation Forest Liu et al (2008) et OneClass Support vector machine OC-SVM Schölkopf et al (2000). Chaque famille de méthodes a ses propres avantages, inconvénients et domaines d'application, et le choix de la méthode dépend souvent du contexte spécifique et des caractéristiques des données à traiter.

Dans cet article, nous appliquons 3 méthodes de détection d'anomalies issues de 3 familles différentes. Il s'agit de $K$-medoids, LOF et IF. $K$-medoids est une technique de clustering basée sur la distance, similaire à $K$-means, mais le centroïde de chaque classe est remplacé par un médoïde qui est l'observation la plus centrale de la classe c'est-à-dire l'observation dont la dissimilarité moyenne avec les autres observations de la classe est minimale. Les médoïdes peuvent être utilisés pour détecter des anomalies en calculant la distance entre chaque observation et le médoïde de sa classe. Si la distance est supérieure à un seuil prédéfini, l'observation est considérée comme une anomalie (Munz et al (2007)).

Le deuxième modèle utilisé dans nos expériences est Local Outlier Factor (LOF) qui est une méthode basée sur la densité. Il s'agit de calculer un indice de densité de chaque point et d'en déduire la densité d'atteignabilité locale d'une observation à comparer avec celles de ses plus proches voisins. Le score LOF de l'observation est alors le rapport entre la densité d'atteignabilité moyenne de ses $K$ voisins les plus proches et sa propre densité. Les anomalies sont les observations ayant un LOF $> 1$. Cela signifie qu'elles sont localisées dans les régions de faible densité.

Isolation Forest (IF) est le troisième modèle utilisé dans nos expériences. C'est une méthode inspirée des forêts aléatoires, utilisée dans la détection d'anomalies. Elle repose sur la construction d'arbres de manière récursive pour isoler les observations. Un score d'anomalie est attribué à chaque observation en sélectionnant aléatoirement une variable et la valeur de coupure. Ce score dépend du nombre de coupures nécessaires pour isoler l'observation, soit la longueur du chemin entre le nœud racine et le nœud terminal. Plus le nombre de coupures est faible plus la probabilité que l'observation soit une anomalie est élevée, et inversement.

# 3 Détection d'anomalies en données mixtes

La plupart des techniques de détection d'anomalies ne sont compatibles qu'avec des données quantitatives. Pour prendre en compte les données catégorielles, une technique couramment utilisée est One hot encoding (OHE). C'est une méthode de pré-traitement de données qui consiste à transformer les modalités en indicatrices binaires en faisant un tableau disjonctif complet pour les variables qualitatives. Cependant, cette technique peut entraîner une augmentation significative de la dimensionnalité des données, surtout si les variables qualitatives ont de nombreuses modalités. Cela peut poser des problèmes de performance et de temps d'exécution pour les algorithmes de détection d'anomalies, en particulier ceux qui impliquent des calculs complexes comme LOF.

Notre approche (Gadacha et al (2023)) consiste à réaliser une analyse factorielle de données mixtes (AFDM) qui permet d'analyser un jeu de données où les observations sont décrites à la fois par des variables quantitatives et par des variables qualitatives. C'est une technique proposé par Pagès (2004) et qui permet d'obtenir des composantes non corrélées qui sont des combinaisons linéaires des variables initiales. Ces composantes sont utilisées, par la suite, pour la détection d'anomalies à l'aide des algorithmes spécifiques aux données quantitatives.

Comme l'AFDM est une généralisation de l'analyse en composantes principales (ACP), nous nous sommes inspirées des travaux de Joliffe (1986) sur l'ACP et le choix des composantes dans la détection d'anomalies. Il a montré que les premières composantes permettent de détecter les valeurs aberrantes qui respectent la structure de corrélation, alors que, les dernières composantes principales permettent de détecter celles qui ne respectent pas la structure de corrélation entre les variables. Afin de vérifier si cette affirmation est également vraie pour l'AFDM, nous avons décidé d'utiliser dans nos expériences des jeux de données contenant à la fois des anomalies qui respectent et d'autres qui violent la structure de corrélation. L'objectif est de comparer les méthodes et les approches de détection afin de déterminer les types d'anomalies identifiés par les premières et dernières composantes de l'AFDM. Cette analyse nous permettra de sélectionner les composantes appropriées pour détecter chaque type d'anomalies.

# 4 Expérimentations

Dans cette section, nous présentons les expériences que nous avons réalisées et les résultats.

## 4.1 Expériences

L'objectif est de simuler des données afin de déterminer les types d'anomalies détectés par chaque modèle. Nous avons simulé $N$=20000 observations et $P$=15 variables dont 10 quantitatives et 5 qualitatives. Nous avons simulé deux classes d'observations normales, chacune contenant 45% des données, ainsi que 10% de valeurs aberrantes. Les variables numériques ont d'abord été créées à l'aide d'une distribution normale multivariée à 10 dimensions. Les

variables catégorielles ont été créées à l'aide de la distribution multinomiale, chaque variable contenant une catégorie rare. La fréquence de la modalité rare sur chacune des variables ne dépasse pas 1.5% de la totalité des observations.

Nous avons généré quatre types d'anomalies : globales, locales, rares et mixtes (Tab. 1). Les anomalies globales et locales sont générées à partir des variables quantitatives, tandis que les anomalies rares sont créées à partir des variables qualitatives et sont normales sur les variables numériques. Les anomalies mixtes sont anormales sur les deux types de variables quantitatives et qualitatives.

- Anomalies globales : nous avons créé trois catégories d'anomalies globales : globales avec des valeurs élevées, globales avec des valeurs basses et globales qui ne respectent pas la structure de corrélation. Ces anomalies sont générées en utilisant la "loi des 4 sigma" qui est une extension de la "loi des 3 sigma". Les anomalies sont situées à quatre écarts-types de la moyenne.

- Anomalies locales : nous avons créé d'abord deux classes composées d'observations normales en décentrant la moitié des observations. Les anomalies locales sont situées dans trois régions différentes : la première région contient des anomalies locales par rapport à la première classe, la deuxième région contient des anomalies locales par rapport à la deuxième classe et la dernière région contient des anomalies locales situées au milieu des deux classes. En disposant les anomalies locales de cette manière, nous simulons différents scénarios où les comportements anormaux peuvent se produire par rapport à différentes situations.

- Anomalies rares : ce sont des anomalies qui présentent au moins une modalité rare et qui sont normales sur les données quantitatives.

- Les anomalies mixtes : ces anomalies peuvent être à la fois globales et rares, ou locales et rares.

| Observations | Catégories | Proportion | Nombre |
|---|---|---|---|
| données normales | | 90.12% | 18024 |
| Anomalies globales | valeurs elevées | 0.80% | 160 |
| | valeurs faibles | 0.80% | 160 |
| | structure de corrélation non respectée | 0.90% | 180 |
| Anomalies locales | | 2.50% | 498 |
| Anomalies rares | | 2.38% | 476 |
| Anomalies mixtes | globales (elevées) et rares | 0.60% | 120 |
| | globales (faibles) and rares | 0.60% | 120 |
| | globales (structure de corrélation non respectée) et rares | 0.60% | 120 |
| | locales et rares | 0.70% | 141 |

Tab. 1 - Caractéristiques des données simulées.

Dans la première expérience, nous évaluons les performances des modèles dans la détection de chaque type d'anomalies séparément. Pour chaque mise en oeuvre, nous nous restreignons aux observations normales et au type d'anomalies concerné. Cela permet de savoir comment chaque modèle se comporte dans la détection des différents types d'anomalies et de comparer leurs performances. Dans la deuxième expérience, nous évaluons les performances des modèles en présence de tous les types d'anomalies sans restriction. Nous utilisons l'ensemble de données comprenant toutes les observations normales ainsi que tous les types d'anomalies. Cette expérience permet d'évaluer la capacité des modèles à détecter les anomalies dans des scénarios plus complexes où plusieurs types d'anomalies peuvent coexister dans les données.

Sur chaque jeu de données, nous avons appliqué le processus de détection d'anomalies décrit ci-après. Tout d'abord, nous avons transformé les données en OHE et nous avons appliqué les 3 algorithmes de détection d'anomalies suivants : IF, LOF et $K$-medoids. Nous avons effectué une validation croisée avec une recherche en grille des meilleurs hyperparamètres et évaluer les performances sur un autre échantillon test.

Par ailleurs, nous avons appliqué l'AFDM aux données d'apprentissage, de sorte que les observations de l'échantillon de validation et de test ne participent pas à la détermination des composantes principales. Nous avons calculé les composantes pour les observations des échantillons de validation et celui de test considérées comme des observations supplémentaires. Nous avons sélectionné les premières composantes principales sur la base du critère de Kaiser-Guttman (Kaiser (1961)), qui consiste à sélectionner les facteurs pour lesquels la valeur propre est supérieure à 1. Nous avons appliqué les 3 algorithmes de détection d'anomalies d'abord aux premières composantes, puis aux dernières composantes et enfin à toutes les composantes de l'AFDM. Pour chaque modèle, nous avons effectué une validation croisée avec recherche en grille des meilleurs hyperparamètres.

## 4.2 Résultats

Le tableau 2 résume les résultats des modèles de détection d'anomalies sur les données après OHE et les données après AFDM pour chaque type d'anomalies.

| Approches | Méthodes | simulation | | | | |
|---|---|---|---|---|---|---|
| | | globale | locale | mixtes | rare | tous types |
| One hot encoding (OHE) | IF | 100% | 100% | 99.99% | 50.95% | 90.92% |
| | LOF | 100% | 80.36% | 99.99% | 51.63% | 90.10% |
| | $K$-medoids | 100% | 100% | 100% | 51.78% | 90.34% |
| AFDM (toutes composantes) | IF | 99.40% | 99.99% | 100% | **53.88%** | 90.34% |
| | LOF | 100% | 100% | 100% | 43.62% | **91.28%** |
| | $K$-medoids | 100% | 76.81% | 100% | 45.06% | 79.60% |
| AFDM (premières composantes) | IF | 99.95% | 87.89% | 100% | 52.44% | 75.08% |
| | LOF | 100% | 99.19% | 99.99% | 45.95% | 87.75% |
| | $K$-medoids | 100% | 87.16% | 100% | 47.63% | 72.36% |
| AFDM (dernières composantes) | IF | 68.70% | 99.99% | 100% | 53.35% | 86.53% |
| | LOF | 84.18% | 99.90% | 100% | 47.27% | 70.03% |
| | $K$-medoids | 67.60% | 67.72% | 100% | 50.38% | 71.83% |

Tab. 2 – Tableau comparatif des performances des 3 modèles (IF, LOF et $K$-medoids) en appliquant les différentes approches pour la détection de chaque type d'anomalies.

Les résultats de nos expériences montrent que les performances des modèles varient en fonction de l'approche utilisée et du type d'anomalies détectés.

1. Les anomalies globales : elles sont bien détectées sur les données encodées, les premières et toutes les composantes de l'AFDM (AUC égale à 1). Néanmoins, elles sont moins bien détectées sur les dernières composantes de l'AFDM, l'AUC varie entre 67.60% et 84.18%).

2. Les anomalies locales : LOF, en tant que modèle dédié à la détection d'anomalies locales, est très efficace sur les premières, les dernières et toutes les composantes de l'AFDM. L'AUC varie entre 99% et 100%. Cependant, il est moins performant sur les données encodées avec un AUC égal à 80.36%. $K$-medoids semble être performant uniquement sur les données OHE. IF est très performant dans la détection de ce type d'anomalies quelque soit l'approche avec de performances légèrement plus faibles sur les premiers composantes de l'AFDM avec un AUC de 87.89%.

3. Les anomalies mixtes : Elles semblent être facilement détectables par les modèles sur les données après OHE et les composantes de l'AFDM.

4. Les anomalies rares: elles sont les plus difficile à détecter. Le meilleur modèle est IF sur toutes les composantes de l'AFDM avec un AUC de 53.88%. Cela peut s'expliquer par le fait que LOF et $K$-medoids reposent sur le calcul des distances entre les observations. Lorsque les données sont qualitatives, le calcul peut être plus complexe par rapport aux données quantitatives. Il nécessite souvent des techniques spécifiques, telles que l'utilisation de mesures de similarité appropriées (ex: les coefficients de Jaccard, Russel-Rao et Sokal-Michener). Cette complexité peut affecter les performances des modèles de détection d'anomalies, en particulier ceux qui reposent fortement sur ces calculs.

5. En cas de données contenant différents types d'anomalies : le meilleur modèle est LOF sur toutes les composantes de l'AFDM avec un AUC de 91.28%.

En comparant les performances des méthodes après l'AFDM, nous notons une différence entre les résultats sur les premières composantes principales et les dernières (à l'exception du cas des anomalies mixtes). Les résultats sur les premières composantes principales sont meilleurs dans la détection des anomalies globales. En revanche, les dernières composantes sont meilleurs dans la détection des anomalies rares et locales, à l'exception de $K$-medoids qui détecte mieux les anomalies sur les premières composantes.

Ces résultats montrent que l'importance des différentes composantes principales : Les premières composantes de l'AFDM expliquent la plus grande partie de la variance totale des données. Ceci les rend efficaces pour détecter les anomalies globales. En revanche, les dernières composantes peuvent contenir des informations plus spécifiques ce qui les rend plus adaptées à la détection d'anomalies rares ou locales.

En synthèse, l'usage de toutes les composantes de l'AFDM permet d'améliorer les performances des modèles en particulier dans la détection des anomalies rares et en présence de tous type d'anomalies.

# 5 Conclusion

La détection d'anomalies est très importante et utile dans de nombreux domaines notamment la finance, la sécurité, la santé etc. Dans cette étude, nous avons présenté une brève revue de certaines méthodes de détection d'anomalies. Un des inconvénients de la plupart des méthodes est qu'elles sont applicables uniquement sur des variables qualitatives et ne parviennent souvent qu'à identifier des anomalies globales et locales.

Dans cette étude, nous avons présenté notre approche basée sur les composantes de l'AFDM et qui vise à améliorer la détection d'anomalies en particulier dans le cas de données mixtes. Nous avons évalué les performances de cette approche comparée à celle des données en OHE dans la détection de 4 types d'anomalies : globales, locales, rares et mixtes. Les résultats de nos expériences ont montré que les modèles sont très efficaces dans la détection des anomalies mixtes. C'est le cas des anomalies globales quelque soit l'approche à l'exception des dernières composantes de l'AFDM. En revanche, les performances dans la détection des anomalies rares sont globalement faibles, l'usage de notre approche a donné un AUC plus élevé que celui des données après OHE. LOF qui est un modèle dédié à la détection des anomalies locales n'a pas été performant sur les données encodées. Néanmoins, il était efficace sur les composantes de l'AFDM. En réalité, les données peuvent contenir diffèrent types d'anomalies. De ce fait, nous avons réalisé la même expérience sur des données contenant à la fois les 4 types d'anomalies. Les résultats ont montré l'intérêt de notre approche dans la détection des anomalies en présence de plusieurs types d'anomalies. Cependant, des évaluations plus formelles, notamment sur de nouvelles simulations où les anomalies sont moins facilement identifiables, sont nécessaires. Nous poursuivons nos travaux dans ce sens.

# Bibliographie


Aggarwal, C. C. (2017). Outlier Analysis (Second Edition ed.). *Springer International Publishing AG.*

Agrawal, S. et Agrawal, J.(2015), Survey on Anomaly Detection using Data Mining Techniques, *Procedia Computer Science*, 708 – 713

Akoglu, L., Tong, H. Vreeken, J. et Faloutsos, C. (2012), Fast and reliable anomaly detection in categorical data. *In CIKM*, pages 415–424.

Angiulli, F. et Fassetti, F. (2009), An efficient algorithm for mining distance-based outliers in very large datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):4:1–57.

Arthur, D. et Vassilvitskii, S. (2007) K-Means++: The Advantages of Careful Seeding Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete algorithms, *Society for Industrial and Applied Mathematics* , Philadelphia, 1027-1035.

Basora, L. , Olive, X., Dubot, D. (2019), Recent Advances in Anomaly Detection Methods Applied to Aviation. Aerospace, *MDPI*, 6 (117), pp.1-27.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). Lof : identifying densitybased local outliers. *In ACM sigmod record* , Volume 29, pp. 93–104. ACM..

Chalapathy, R. et S. Chawla (2019). Deep learning for anomaly detection : A survey, arXiv preprint arXiv:1901.03407.

Chandola, V., Banerjee, A. et Kumar, V. (2009). Anomaly detection : A survey. *ACM computing surveys* 41(3), 15.

Ester, M., Kriegel, H.P., Sander, J. et Xu, X. (1996), A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proceedings of the Second KDD'96 International Conference on Knowledge Discovery and Data Mining* , Portland, OR, USA, pp. 226–231.

Gadacha, H., Kubicki, P., Niang, N. (2023). Détection d'anomalies en présence de données quantitatives et qualitatives. *SFdS2023*, [consulté le 19 février 2024]. Disponible sur : https://drive.google.com/file/d/14qQcZ69m4O3Ez-eunTgzLQzl8CjtpJDF/view

Govaert, G. (2003), Analyse des données, *Lavoisier*, Paris.

Joliffe, I. T. (1986), Principal Component Analysis. *Springer Verlag.*

Kaiser, F. (1961). A note on Guttman's lower bound for the number of common factors, *British Journal of Statistical Psychology*, 14, 1-2.

Knorr, E. M. , Ng, R. T. et Tucakov, V. (1999), Distance-based outliers: algorithms and applications, *The VLDB Journal*, 8(3):237–253.

Koufakou, A., Ortiz, E., Georgiopoulos, M., Anagnostopoulos, G., and Reynolds, K. (2007). A scalable and efficientoutlier detection strategy for categorical data. *In Proceedings of the IEEE International Conference on Tools with ArtificialIntelligence ICTAI*, pages 210–217,

Patras-Peloponnese-Greece.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2008). Isolation forest. *In 2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE.

Munz,G., Li, S. et Carle, G. (2007), Traffic Anomaly Detection Using K-Means Clustering, *GI/ITG Workshop MMBnet*, p.1-8

Orair, G. H., Teixeira, C. H., Wang, Y. et Meira Jr, T. W. (2010), Distance-Based Outlier Detection: Consolidation and Renewed Bearing, *Proceedings of the VLDB Endowment*, Volume 3, p 1469–1480.

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée*, 5–37.

Rokhman, N., Subanar, and Winarko, E. (2016). Improving the performance of outlier detection methods for categorical data by using weighting function. *Journal of Theoretical and Applied Information Technology*, 83:327–336.

Schölkopf, B., R. C. Williamson, A. J. Smola, J. Shawe-Taylor, et J. C. Platt (2000), Support vector method for novelty detection *In Advances in neural information processing systems*, pp. 582–588.

Shyu, M., Sarinnapakorn, K., Kuruppu-Appuhamilage, I., Chen, S., Chang, L. W., et Goldring, T. (2005). Handling nominal features in anomaly intrusion detection problems. *In Proceedings of the International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications*, pages 55–62, Tokyo, Japan.

Syarif I., Prugel-Bennett A. et Wills G. (2012) , Data mining approaches for network intrusion detection from dimensionality reduction to misuse and anomaly detection, *Journal of Information Technology Review*, p. 70-83.

Taha, A. and Hadi, A. S. (2016). Pair-wise association for categorical and mixed attributes. *Information Sciences*, 346:73–89.

Togbe, M., Chabchoub, Y., Boly, A. et Chiky, R. (2020), Étude comparative des méthodes de détection d'anomalies. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances EGC, vol. RNTI-E-36, pp.109-120.

Velmurugan, T. et Santhanam, T. (2010), Computational Complexity between k-Means and k-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points, *Journal of Computer Science*, 6 (3): 363-368.

Wang, Y. (2008). Statistical Techniques for Network Security: Modern Statistically-Based Intrusion Detection and Protection.*IGI Global*, New York, NY, USA.

Wibisono, S., Anwar, M.T , Supriyanto, A., et Amin, I.H.A (2020), Multivariate weather anomaly detection using DBSCAN clustering algorithm, *Journal of Physics: Conference Series* , Volume 1869.

Xu, X., Liu, H. , et Yao, M. (2019), Recent Progress of Anomaly Detection, *Hindawi Complexity*, pages 1-11, January.

# Détection non supervisée d'anomalies dans les images satellites pour le monitoring des surfaces océaniques à l'aide de l'ACP robuste et du test de Goodness of Fit basé sur la distance de Wasserstein entre processus ponctuels

Julien Bastian [1] & Stéphane Chrétien [1,2] & Ben Gao [1,3] & Rémi Vaucher[1,3]

[1] *Laboratoire ERIC, France. E-mail : julien.bastian, stephane.chretien, ben.gao, remi.vaucher@univ-lyon2.fr*
[2] *The Alan Turing Institute, London*
[3] *HALIAS Technologies, France. E-mail : ben.gao, remi.vaucher@halias.fr*

**Résumé.** L'utilisation de l'imagerie satellitaire pour la détection d'anomalies à la surface de la mer offre une solution prometteuse pour une surveillance systématique. Nous proposons une approche efficace et non supervisée pour la détection de certaines anomalies de la surface de la mer à l'aide d'une ACP robuste, combinée à un test d'adéquation utilisant la distance de Wasserstein pour caractériser statistiquement la différence entre la forme des anomalies et la loi uniforme.

**Mots-clés.** Robust PCA, Détection d'anomalie, Monitoring de la surface marine.

**Abstract.** The use of satellite imagery for sea surface anomaly detection offers a promising solution for systematic monitoring. We propose an efficient and unsupervised approach for the detection of certain sea surface anomalies using Robust PCA, combined with a goodness-of-fit test using Wasserstein distance to statistically characterize the difference between the shape of the anomalies and the uniform law.

**Keywords.** Robust PCA, Anomaly detection, Marine surface monitoring.

# 1 Présentation du problème de détection

La détection de fuites d'hydrocarbures dans les environnements marins est un défi majeur pour l'industrie pétrolière et les organismes de surveillance environnementale. L'utilisation d'images satellite pour la détection offre une alternative prometteuse en fournissant une couverture étendue et une actualisation régulière des données observées.

## 1.1 Décomposition des images par Robust PCA

L'analyse de ces images par des méthodes avancées de statistique en haute dimension, utilisant la sparsité et la sparsité spectrale, telle la Robust PCA, permet d'identifier les

signaux faibles associés aux fuites d'hydrocarbures, qui seraient autrement noyés dans le bruit de fond ou masqués par d'autres caractéristiques de la surface marine.

En effet la surface marine dispose d'une régularité due au mouvement naturel de l'eau, qui peut être représenté par une matrice de rang faible.

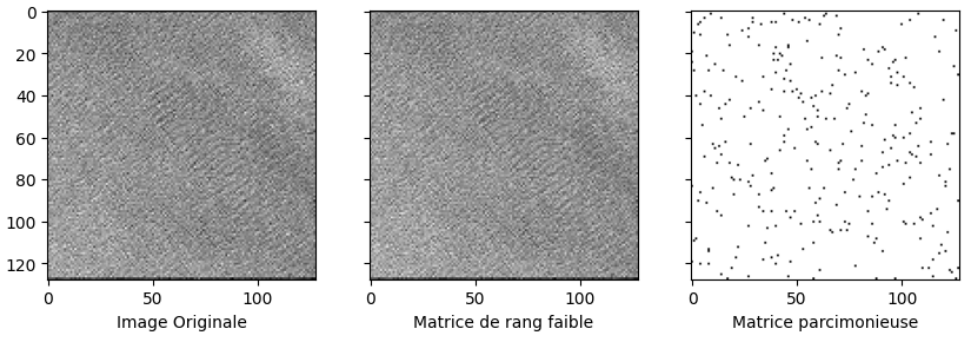
### 1.1.1 Rappels sur l'ACP robuste

L'ACP robuste est une technique puissante pour séparer une matrice d'observations en composantes de structure différente :

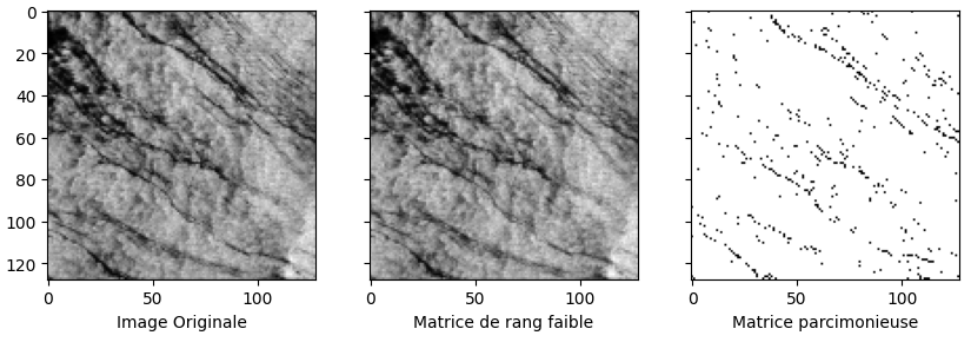
$$X = L + S,$$

où $L$ est une matrice de rank faible et $S$ une matrice parcimonieuse.

Appliquée à l'analyse d'images satellite pour la détection de fuites d'hydrocarbures, cette méthode consiste à décomposer la matrice image en la somme de deux matrices distinctes : une matrice de rang faible et une matrice sparse (ou parcimonieuse). La matrice de rang faible capture la partie stationnaire ou de fond de l'image, incluant les caractéristiques constantes de la surface de l'eau et les objets stables, tandis que la matrice sparse isole les anomalies ou les éléments transitoires, comme les fuites d'hydrocarbures.




(a) Sans anomalie




(b) Avec anomalie

FIGURE 1 – Comparaison des décomposition Rang faible/Sparse pour des images sans et avec anomalies

### 1.1.2 Méthode numérique

Plusieurs approches numériques ont été développées pour l'identification des deux matrices $L$ et $S$ dans la décomposition de l'image.

**Moindres carrés pénalisés par la norme nucléaire et la norme $\ell_1$.** L'objectif est de trouver une matrice qui minimise la somme d'une fonction de perte (habituellement les moindres carrés) et d'une pénalité combinaison linéaire de la norme nucléaire de la matrice $L$ et de la norme $\ell_1$ de la matrice $S$. La norme nucléaire de la matrice $L$, notée $\|L\|_*$ est la somme des valeurs singulières de la matrice. Le problème d'estimation de $L$ et $S$ revient alors à résoudre le problème d'optimisation

$$\min_{L,S\in\mathbb{R}^{d_1\times d_2}} \frac{1}{2}\|X-L-S\|_F^2+\lambda\|L\|_*+\mu\|S\|_1, \tag{1}$$

pour lequel une selection précise des paramètres de régularisation est requise et peut être fondée sur la Validation Croisée, par exemple. Cette approche étend celle proposée dans [Candes et al., 2009] au cas bruité.

**Méthode de minimisation alternée.** Une méthode de minimisation alternée rapide peut être mise en œuvre. Il est possible par exemple d'implémenter la méthode proposée dans [Cai et al., 2019]. Nous nous sommes restreint à la méthode Principal Component Pursuit de [Candes et al., 2009], dans les expériences proposées.

# 2 Détection par test de Wasserstein Goodness of Fit

Dans le contexte de la détection de fuites, les composantes non-nulles de la matrice Sparse correspondent aux pixels d'intérêt où les fuites sont susceptibles d'être présentes. Le problème principal est que les pixels sélectionnés ne permettent pas de circonscrire une zone bien définie mais se répartissent plutôt comme la réalisation d'un processus ponctuel dont le support spatial coïncide avec les zones contaminées par les hydrocarbures.

L'approche proposée pour la détection consiste à verfier à l'aide d'un test statistique que le support du processus ponctuel révélé par la matrice $S$ est différent de celui d'un processus de Poisson associé à une intensité uniforme sur l'image.

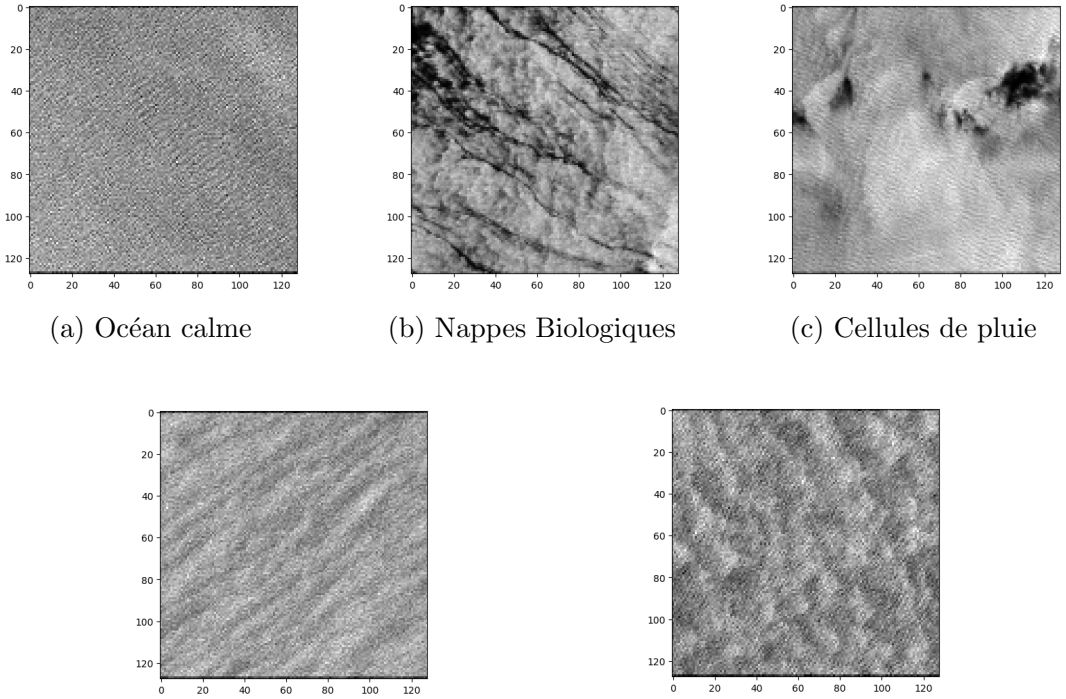
## 2.1 Mise en place du test

Pour valider l'efficacité de cette méthode dans l'identification des fuites d'hydrocarbures, un test de goodness of fit est souvent réalisé en utilisant la distance de Wasserstein. Cette distance mesure l'écart entre la distribution spatiale des composantes identifiées comme fuites d'hydrocarbures et une distribution de référence pour laquelle aucune hypothèse sur la forme

du support n'est faite. Un écart significatif indique que la matrice sparse $S$ détecte efficacement les anomalies correspondant aux fuites d'hydrocarbures, validant ainsi la pertinence de la décomposition obtenue par Robust PCA.

# 3 Expériences

## 3.1 Données



(a) Océan calme (b) Nappes Biologiques (c) Cellules de pluie

(d) Traînées de vent (e) Micro cellules convectives

FIGURE 2 – Illustratins des différentes classes considérées.

Les données que nous avons utilisées pour les expériences numériques permettant d'évaluer les performances de l'approche sont celles proposées dans [Wang et al., 2019]. Ces données présentent une collection d'images satellites de type SAR, contenant divers types de configurations de surfaces marines labelisées suivant le type de phénomène présent. Les classes représentées sont "Biological Slicks", "MicroConvective Cells", "Pure Ocean Waves", "Rain Cells" et "Wind Streaks", illustrées en Figure 2.

## 3.2 Résultats

Nous avons choisi l'hyperparamètre $\lambda = \frac{2}{\sqrt{n_{max}}}$ pour la Robust PCA parmi les options suivantes : $\frac{1}{\sqrt{n_{max}}}$, $\frac{2}{\sqrt{n_{max}}}$ et $\frac{3}{\sqrt{n_{max}}}$. Seules les entrées négatives de la matrice parcimonieuse

| Prédiction | Océan Pure | Nappes biologiques | Cellules de pluie | Trainées de vent | Micro cellules convectives |
|---|---|---|---|---|---|
| Rejet | 0.060 | 0.895 | 0.730 | 0.115 | 0.125 |
| Non rejet | 0.940 | 0.105 | 0.270 | 0.885 | 0.875 |

TABLE 1 – Taux d'image pour lesquelles le test de Wasserstein Goodness of Fit rejette l'uniformité avec le seuil de significativité $\alpha = 0.05$.

obtenue sont conservées, les anomalies considérées ici correspondant à des zones de luminosité basses. Parce que celles-ci représentent les nappes que nous souhaitons détecter. De plus en supposant que les entrées non nulles proviennnet d'une réalisation d'un processus ponctuel, nous préservons la structure du support de la matrice $S$ plutôt que l'amplitude de ses entrées. En ce qui concerne le test de Wasserstein Goodness of Fit, nous avons généré, pour chaque image testé, 90 échantillons suivant une loi uniforme bi-dimensionnelle et qui ont la même taille que le nombre de points dans $S$, afin d'obtenir la distribution empirique de $T_n : W_2^2(\hat{P}_n, P_{uniform})$.

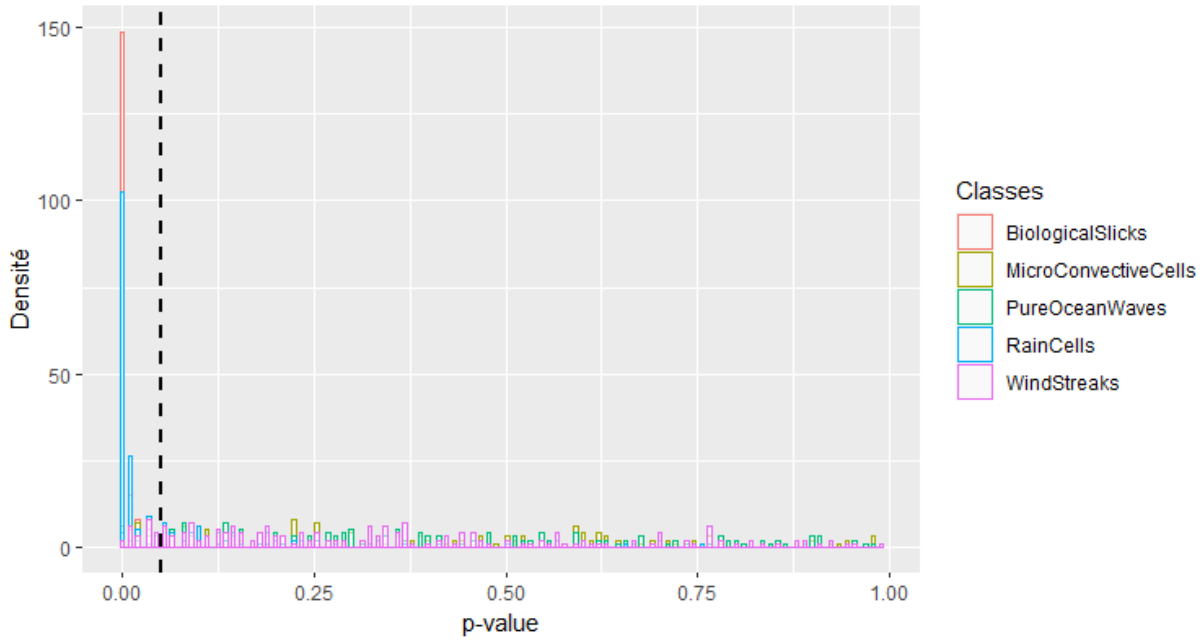



FIGURE 3 – Distribution des p-values du test de Wasserstein Goodness of Fit, la ligne verticale représente le seuil de significativité $\alpha = 0.05$

Les Figures 3 et 4 montrent que les deux classes principalement détectées par notre méthode sont les classes Biological Slicks et Rain Cells. Les matrices Sparses de la décomposition des images de type Rain Cells ont une forme caractéristique de type réalisation d'une gaussienne bi-variée et sont facilement détectables en comparant la distribution à une gaussienne plutôt qu'a un processus ponctuel uniforme. Les classes "MicroConvective Cells", "Pure Ocean Waves" et "Wind Streaks" ont une distribution beaucoup plus proche de la distribution uniforme. Cela démontre que ces anomalies sont d'un type pulvérisé et nous intéressent moins en terme de monitoring de la surface marine. Le tableau 1 illustre les performances de

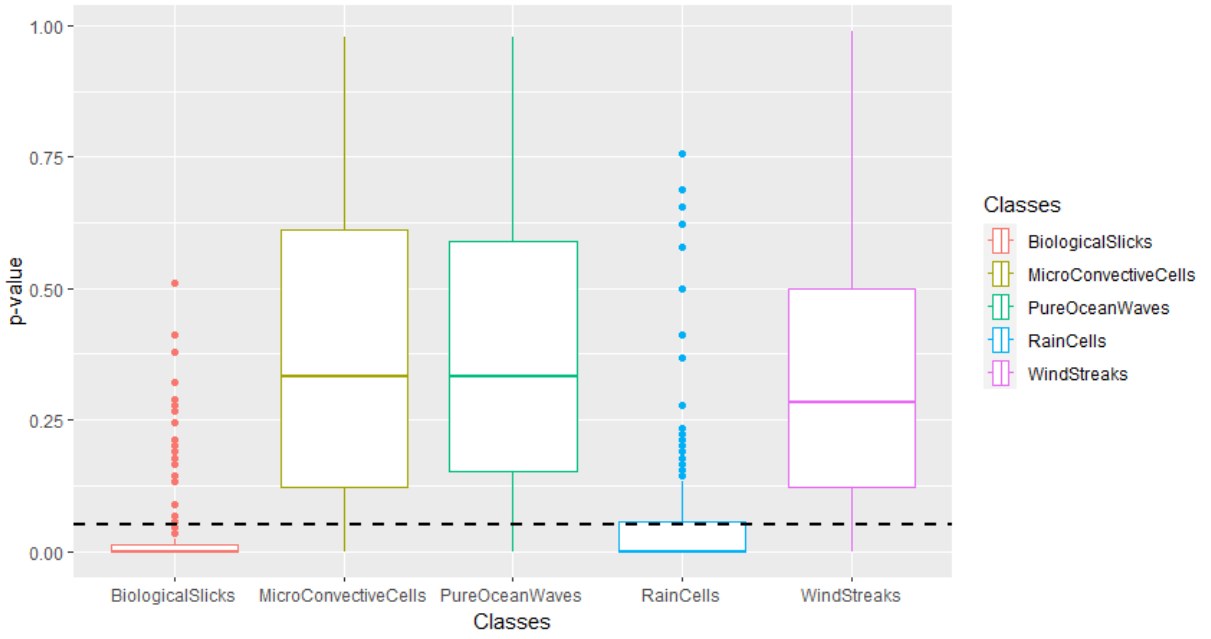



FIGURE 4 – Boxplot des p-values pour le test d'adequation fondé sur les distances de Wasserstein, la ligne horizontale représente le seuil de significativité $\alpha = 0.05$.

la méthode vis à vis de chaque classe considérée.

# 4 Conclusion

En conclusion, l'application de la Robust PCA à l'analyse d'images satellite pour la détection de fuites d'hydrocarbures représente une avancée significative dans la surveillance environnementale des installations pétrolières.

## 4.1 Avantages de l'approche proposée

La méthode Robust PCA combinée avec le test sur la distance de Wasserstein pour l'analyse d'images satellite en vue de la détection de fuites offre plusieurs avantages.

— Premièrement, elle permet une discrimination fine entre les caractéristiques de fond et les anomalies, rendant possible la détection de fuites même dans des conditions où les signaux sont faibles ou masqués par d'autres éléments de l'image.
— L'approche proposées est complétement non-supervisée et fournit un cadre automatisé qui peut être régulièrement mis à jour avec de nouvelles données d'observation.
— L'utilisation de la distance de Wasserstein comme critère de validation ajoute une couche supplémentaire de fiabilité à la détection, en offrant une mesure quantitative de la performance de la méthode.

## 4.2 Perspectives

Une amélioration de la puissance du test pourra être proposée dans une extension prochaine de ces travaux, en incorporant des *a priori* plus spécifiques sur la forme de la structure du support du processus ponctuel à détecter.

# Bibliographie

[Cai et al., 2019] Cai, H., Cai, J.-F., and Wei, K. (2019). Accelerated alternating projections for robust principal component analysis. *The Journal of Machine Learning Research*, 20(1) :685–717.

[Candes et al., 2009] Candes, E. J., Li, X., Ma, Y., and Wright, J. (2009). Robust principal component analysis ?

[Wang et al., 2019] Wang, C., Mouche, A., Tandeo, P., Stopa, J. E., Longépé, N., Erhard, G., Foster, R. C., Vandemark, D., and Chapron, B. (2019). A labelled ocean sar imagery dataset of ten geophysical phenomena from sentinel-1 wave mode. *Geoscience Data Journal*, 6(2) :105–115.

# Régression

# Test de rupture de régression

Zaher Mohdeb

*Faculté de Génie des Procédés*
*Université de Constantine 3, Salah Boubnider*
*Constantine, Algérie*
*et*
*Laboratoire de Mathématiques et Sciences de la Décision*
*Université frères Mentouri*
*Constantine, Algérie*
*E-mail: zaher.mohdeb@univ-constantine3.dz*

**Résumé.** On considère un modèle de régression non paramétrique à erreurs homoscédastiques et un échantillonnage fixé; notre objectif est de construire le test de l'hypothèse linéaire contre les alternatives de ruptures de modèle et ce sans condition de régularité sur la fonction de régression aussi bien sous l'hypothèse nulle que sous l'alternative, ce qui inclut la possibilité que les fonctions soient höldériennes d'ordre supérieur à 1/2. On établit la normalité asymptotique de la statistique de test sous l'hypothèse nulle ainsi que sous l'hypothèse alternative de ruptures de modèle.

**Mots-clés.** Hypothèse linéaire, régression non paramétrique, rupture de modèle.

**Abstract.** We consider a regression model in the case of a homoscedastic error structure and fixed design, our aim is to build the test of the linear hypothesis versus regime switching models without regularity condition, and also under either the null or the alternative hypotheses, which includes the possibility of functions satisfying the Hölder condition of order greater than 1/2. We establish the asymptotic normality of the test statistic under the null hypothesis and the alternative one.

**Keywords.** Linear hypothesis, nonparametric regression, regime switching.

# 1 Introduction

On considère le modèle de régression suivant

$$Y_{i,n} = f(t_{i,n}) + \varepsilon_{i,n}\,, \quad i = 1, \ldots, n\,, \tag{1}$$

où $f$ est une fonction réelle inconnue, définie sur l'intervalle $[0,1]$ et $t_{i,n}$, $i = 1, \ldots, n$, est un échantillonnage fixé de l'intervalle $[0,1]$. Les erreurs $\varepsilon_{i,n}$ forment un tableau triangulaire de variables aléatoires d'espérance nulle et de variance finie $\sigma^2$.

Soient $g_1, \ldots, g_p$ des fonctions définies sur $[0,1]$ et linéairement indépendantes et soit $E_p$ l'espace vectoriel engendré par $g_1, \ldots, g_p$. On veut tester l'hypothèse:

$$H_0 : \ f \in E_p \quad \text{contre} \quad H_1 : \begin{cases} \exists\, s \in ]0,1[ \ \text{ tel que } \ f = \phi \mathbb{1}_{[0,s]} + \psi \mathbb{1}_{]s,1]}, \\ \phi \in E_p\,, \ \ \psi \text{ Riemann intégrable et } f \notin E_p\,. \end{cases} \tag{2}$$

La plupart des travaux sur les tests d'hypothèses dans le modèle (1) supposent des conditions de régularité sur $f$, $g_1, \ldots, g_p$ ; généralement ces fonctions sont supposées höldériennes. On peut citer, sans être exhaustif, Cox et al (1988), Eubank et Spiegelmann (1990), Eubank et Hart (1992), Azzalini et Bowman (1993), Härdle et Mammen (1993). Les tests basés sur l'estimation sur la $L^2$-distance entre $f$ et $E_p$ sont étudiés par Dette et Munk ((1998), Munk et Dette (1998), Mohdeb et Mokkadem (2004), avec l'hypothèse que $f$ est höldérienne d'ordre $\gamma > 1/2$.

Dans ce travail, on applique l'approche utilisée dans Mohdeb et Mokkadem (2015) et Lessak et Mohdeb (2015) pour construire le test d'hypthèses (2) dans le modèle (1). On suppose que $f$, $g_1, \ldots, g_p$ sont Riemann-intégrables; sous cette seule condition sur les fonctions, on établit la normalité asymptotique de la statistique de test qui permet de construire le test (2) et d'avoir la puissance pour des alternatives de ruptures de modèle.

Dans la section 2, on introduit les hypothèses et on présente notre résultat principal. Des simulations ont été menées pour étudier la performance du test proposée dans la section 3.

# 2 Hypothèses et résultats

On considère le modèle de régression (1) et $E_p$ est l'espace vectoriel engendré par des fonctions fixées $g_1, \ldots, g_p$ définies sur $[0,1]$ et linéairement indépendantes.

Nos hypothèses sont les suivantes:

- (A1) $\max\limits_{i=2,\ldots,n} \left|(t_{i,n} - t_{i-1,n}) - \dfrac{1}{n}\right| = o\left(\dfrac{1}{n}\right)$ ;
- (A2) $\forall n$, $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ sont indépendantes et $\exists C \in \mathbb{R}^+$ tel que $E(\varepsilon_{i,n}^4) < C, \ \ \forall i, n$ ;
- (A3) Les fonctions $f$, $g_1, \ldots, g_p$ sont Riemann-intégrables.

**Remarque.** Notons que l'hypothèse (A1) implique que pour toute fonction $h$ Riemann-intégrable

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(t_{i,n}) = \int_0^1 h(t)\, dt.$$

De plus, nous supposons que les fonctions que nous considérons appartiennent également

à $L^2(dt)$, muni de son produit scalaire usuel. On pose,

$$\mathcal{D}^2(f) := \min_{v \in E_p} \|f - v\|^2$$

la distance entre $f$ et le sous-espace $E_p$, $Y := (Y_{1,n}, \ldots, Y_{n,n})'$, $f_n := \Big(f(t_{1,n}), \ldots, f(t_{n,n})\Big)'$, $g_{k,n} := \Big(g_k(t_{1,n}), \ldots, g_k(t_{n,n})\Big)'$, $k = 1, \ldots, p$, et $G := (g_{1,n}, \ldots, g_{p,n})$.

On note aussi $E_{p,n}$, le sous-espace de $\mathbb{R}^n$ engendré par $\{g_{1,n}, \ldots, g_{p,n}\}$ qui est une discrétisation du sous-espace $E_p$, $\Pi_n = G(G'G)^{-1}G'$, la matrice de projection sur $E_{p,n}$ et $\Pi_n^{\perp} = I_n - G(G'G)^{-1}G'$, la matrice de projection sur l'espace orthogonal de $E_{p,n}$, où $I_n$ est la matrice identité $n \times n$.

On considère la statistique suivante définie par

$$D_n^2 := \frac{1}{n} Y' \Pi_n^{\perp} Y .$$

On vérifie que

$$E(D_n^2) = \widetilde{D}_n^2 + \frac{n-p}{n}\sigma^2 , \quad \text{où} \quad \widetilde{D}_n^2 = \frac{1}{n} f_n' \Pi_n^{\perp} f_n .$$

On est amené ainsi à considérer $D_n^2 - \dfrac{n-p}{n}\sigma^2$, mais $\sigma^2$ est inconnu. On l'estime à l'aide de l'estimateur suivant, introduit par Gasser, Sroka, et Jennen-Steinmetz (1986)

$$S_\varepsilon^2 = \frac{1}{6(n-2)} \sum_{i=2}^{n-1} \left(Y_{i+1,n} + Y_{i-1,n} - 2Y_{i,n}\right)^2 .$$

On obtient ainsi la statistique de test donnée par

$$\widehat{D}_n^2 = D_n^2 - \frac{n-p}{n} S_\varepsilon^2 ;$$

et on rejette l'hypothèse $\mathrm{H}_0$ : "$f \in E_p$" si $\widehat{D}_n^2 > u_\alpha$, où $u_\alpha$ est un nombre réel positif.

Notre résultat principal est le suivant.

**Théorème 1** *Si les conditions* $(A1)$, $(A2)$ *et* $(A3)$ *sont satisfaites, alors*

$$\sqrt{n}\left\{\widehat{D}_n^2 - \widetilde{D}_n^2 + B_n(f)\right\} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{17}{9}\sigma^4 + 4\sigma^2\mathcal{D}^2(f)\right) ,$$

*où* $B_n(f) = \dfrac{1}{6n} \displaystyle\sum_{i=2}^{n-1} \Big(f(t_{i+1,n}) + f(t_{i-1,n}) - 2f(t_{i,n})\Big)^2$.

**Remarque.** Sous $\mathrm{H}_0$, on a $\widetilde{D}_n^2 = 0$ mais $B_n(f)$ n'est pas nécessairement nul, ni même négligeable en général.

# 3 Applications

## 3.1 Test dans un modèle de régression localement höldérienne

On suppose que, dans le modèle de régression étudié, $f$ est une fonction localement höldérienne d'ordre inconnu (ou seulement Riemann-intégrable) et on considère un espace vectoriel $E_p$ tel que

- (A4) les fonctions $g_1, \ldots, g_p$ sont localement höldériennes d'ordre $\gamma > 1/2$.

Dans ce cas, il existe une subdivision de $[0,1]$ en intervalles $I_1, \ldots, I_q$ et un réel $\delta > 1/2$, ($\delta$ est le plus petit des ordres des fonctions $g_1, \ldots, g_p$) tel que tout élément de $E_p$ est Hölder d'ordre $\delta$ sur chaque $I_j$. Cela n'est évidemment pas le cas pour l'alternative, c'est-à-dire l'ensemble des fonctions localement höldériennes qui n'appartiennent pas $E_p$.

Il est facile de vérifier que pour tout $f \in E_p$,

$$B_n(f) = o\left(\frac{1}{\sqrt{n}}\right).$$

On a donc la proposition.

**Proposition 1** *Si les conditions* $(A1)$, $(A2)$ *et* $(A4)$ *sont satisfaites on a, sous* $H_0$,

$$\sqrt{n}\widehat{D}_n^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{17}{9}\sigma^4\right).$$

Cette proposition donne le niveau asymptotique du test; le Théorème 1 donne la puissance pour des alternatives qui peuvent être höldériennes d'ordre $\delta < 1/2$, (ou seulement Riemann-intégrable). En pratique la variance $\sigma^2$ des erreurs est généralement inconnue, on peut considérer un estimateur consistent $\widehat{\sigma}^2$ de $\sigma^2$. On rejette l'hypothèse nulle $H_0$ : "$f \in E_p$", si

$$\frac{\sqrt{n}}{\widehat{\sigma}^2}\,\widehat{D}_n^2 \;>\; z_{1-\alpha}\,,$$

où $z_{1-\alpha}$ est le $(1-\alpha)$quantile d'une loi normale standard.

## 3.2 Test de rupture de modèle

Soit $E_p$ un sous-espace vectoriel engendré par des fonctions $g_1, \ldots, g_p$ localement höldériennes d'ordre $\gamma > 1/2$. On veut tester l'hypothèse:

$$\mathrm{H}_0 : \quad f \in E_p \quad \text{contre} \quad \mathrm{H}_1 : \quad \begin{cases} \exists\, s \in ]0,1[ \text{ tel que } f = \phi \mathbb{1}_{[0,s]} + \psi \mathbb{1}_{]s,1]} \\ \\ \phi \in E_p\,, \;\; \psi \text{ Riemann-intégrable et } f \notin E_p\,. \end{cases}$$

Comme l'alternative $H_1$ est contenue dans l'alternative du test de la sous-section 3.1, on peut utiliser ce dernier pour tester la rupture.

# 4 Simulations

On a réalisé des simulations pour étudier la puissance au niveau $\alpha = 5\%$ du test de l'hypothèse

$$H_0 : \quad f(t) = t \quad \text{contre} \quad H_1 : f(t) = t\mathbb{1}_{[0,s]}(t) + \beta t\mathbb{1}_{]s,1]}(t)\,, \quad \beta \neq 1\,.$$

On a considéré un échantillonnage régulier, $t_{i,n} = \frac{i-1}{n-1},\ i = 1, \ldots, n;$ avec $n = 64$.

L'hypothèse $H_0$ est rejetée si

$$\left(\frac{9n}{17}\right)^{1/2} \frac{\widehat{D}_n^2}{\widehat{\sigma}^2} > 1.65\,,$$

avec

$$\widehat{D}_n^2 = \frac{1}{n}\sum_{i=1}^{n} |Y_{i,n} - \widehat{a}t_{i,n}|^2 - \frac{n-1}{n}S_\varepsilon^2$$

et

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_{i,n} - \widehat{a}t_{i,n})^2,$$

où

$$\widehat{a} = \frac{\sum_{i=1}^{n} t_{i,n}Y_{i,n}}{\sum_{i=1}^{n} t_{i,n}^2}\,.$$

Pour $s$ et $\beta$, on a considéré les valeurs $s = 0.238$, 0.492, 0.619, 1.000 et $\beta = 0.00$, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00.

On a aussi considéré plusieurs valeurs de la variance $\sigma^2$ du bruit: $\sigma = 0.05$, 0.10, 0.20 et 0.50.

L'analyse des résultats obtenus montre que pour des petites variances de $\varepsilon$, l'hypothèse $H_0$ est rejetée avec une proportion proche de 1.

# Bibliographie

[1] Azzalini, A. and Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc. Ser. B*, **55**, 549-557.
[2] Cox, D., Koh, G., Wahba, G. and Yandell, B. S. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.*, **16** 113-119.

[3] Dette, H., and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 2, 778-800.
[4] Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Stat.*, **20**, 3, 1412-1425.
[5] Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Assoc.*, **85**, 410, 387-392.
[6] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika,* **73**, 625-633.
[7] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, **21**, 4, 1926-1947.
[8] Lessak, R. and Mohdeb, Z. (2015). Testing the linear regression model null hypothesis versus regime switching alternatives. *Afr. Stat.*, **10**, 807-813.
[9] Mohdeb, Z. and Mokkadem, A. (2004). Average squared residuals approach for testing linear hypothesis in nonparametric regression. *J. Nonparametric Stat.*, **16**, 1-2, 3-12.
[10] Mohdeb, Z. and Mokkadem, A. (2015). Testing linear regression models in non regular case. *Comm. Statist. Theory Methods*, **44**, 21, 4476-4490.
[11] Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *Ann. Stat.*, **26**, 6, 2339-2368.

# Reduced run-time and memory complexity regression with a Gaussian processes prior

Amal Omrani [1] & Anis Fradi [2] & Chafik Samir [3]

[1] *Paris Dauphine University, France, amal.omrani@dauphine.psl.eu*
[2] *University of Clermont Auvergne, France, anis.fradi@uca.fr*
[3] *University of Clermont Auvergne, France, chafik.samir@uca.fr*

**Résumé.** Les processus gaussiens ont connu un grand succès d'utilisation et de performance dans plusieurs applications d'apprentissage automatique. Entre autres, ils possèdent des propriétés théoriques et pratiques qui en font une solution flexible et adaptable pour des problèmes de régression et de classification. Cependant, ils présentent des inconvénients limitants en raison de la complexité computationnelle et en mémoire. Dans ce papier, nous proposons une nouvelle méthode pour construire des processus gaussiens efficaces sur le plan computationnel dans le cadre d'une régression ou d'une classification. En particulier, nous démontrons que les modèles proposés ont une complexité en temps et en mémoire inférieure à celle des processus gaussiens standards. Nous confirmons cette meilleure performance, en pratique, grâce à des expériences et des comparaisons variées.

**Mots-clés.** processus gaussiens, régression, classification, regroupement, complexité.

**Abstract.** Gaussian processes have been successfully used in machine learning applications. They have nice theoretical and practical properties that make them a flexible solution for regression and classification problems. However, they have serious limitations due to the memory and computational complexity. In this paper, we propose a new method to build computationally efficient Gaussian processes for regression and classification problems. The proposed models are shown to have lower time and memory complexity than standard Gaussian process models. Furthermore, we confirm this better performance in practice with varied experiences and comparisons.

**Keywords.** Gaussian processes, regression, classification, clustering , complexity.

## 1 Introduction

Gaussian processes (GPs) are powerful probabilistic models used in machine learning and statistics [5]. They define a distribution over functions, where any finite subset of function values follows a multivariate Gaussian distribution. A GP is characterized by a mean function and a covariance function that encapsulates the smoothness and correlations in the function space. This flexibility allows GPs to model a wide range of complex and non-linear relationships, making them particularly valuable in regression, classification, and uncertainty quantification tasks. GPs are flexible statistical models that have gained significant popularity in the fields of econometrics, shape analysis, signal processing, data science and machine

learning [3]. Their ability to provide uncertainty estimates alongside predictions makes GPs essential in decision-making processes where understanding the confidence in the model's output is crucial.

GPs have some limitations, especially when dealing with high number of observations. The two main challenges are complexity and memory usage:

- **Computational Complexity**: The computational complexity of GPs is cubic with respect to the number of data points $N$, making them impractical for large datasets. Inference and learning involve inverting a covariance matrix, which requires $\mathcal{O}(N^3)$ operations. This makes GPs slow and resource-intensive as the dataset size increases.
- **Memory Consumption**: GPs store and manipulate a covariance matrix that grows rapidly with the number of data points $N$, requiring a significant amount of memory. The covariance matrix is of size $\mathcal{O}(N^2)$, leading to high memory consumption for large datasets. This can be a significant bottleneck in memory-limited environments.

Efforts to address these limitations include approximate inference methods like variational approaches and sparse approximations [1]. These methods attempt to reduce the computational and memory complexity of GPs while still providing reasonable performance. In addition, [6] addressed the computational challenges by employing an approximation method for the covariance matrix, a crucial component in computations involving kernel methods. Moreover, [4] introduced an innovative approach using the Fast Fourier Transform (FFT) for stationary covariances. This method efficiently computes and manipulates covariance functions in the frequency domain.

In pursuit of the same objective, we introduce two clustering-based approaches for selecting the most informative observations, focusing on managing large datasets. Within each cluster, we choose representative or centroid points to serve as observations for the GP. This strategy effectively alleviates the computational and memory demands associated with handling the entire dataset, rendering the GP more scalable [2]. By carefully selecting representative points from the clusters, we uphold a strong approximation of the underlying data structure and variability. This enables us to create effective models and accurate predictions while significantly enhancing computational efficiency. This technique strikes a critical balance between model performance and computational feasibility, making GPs a more viable option for larger datasets.

**Organization.** The remainder of this paper is structured as follows. In Section 2, we introduce the standard Gaussian process regression model. Section 3 outlines our proposed method. We provide an illustrative example in Section 4, and conclude the paper in Section 5.

## 2 Standard Gaussian process regression

Gaussian process regression (GPR) main goal is to model and predict the relationship between input data points $\mathbf{x}_i \in \mathbb{R}^d$ and their corresponding output values $y_i \in \mathbb{R}$ from the nonlinear

relationship

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \text{with} \quad \epsilon_i \sim \mathcal{N}(0, \sigma_N^2) \tag{1}$$

where $f(\mathbf{x})$ is assumed to be a realization of a GP $f$. The stochastic process $f$ is characterized by a mean function $m(\mathbf{x})$ and a covariance function (or kernel) $k(\mathbf{x}, \mathbf{x}')$ denoted as follows

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2}$$

The mean $m$ is usually assumed to be zero for which the GP is assumed to be a centered GP. The kernel $k$ typically depends on a set of unknown hyperparameters that need to be learned from the data when maximizing the log marginal likelihood which is proportional to

$$L = -\frac{1}{2}\log|\mathbf{K} + \sigma_N^2\mathbf{I}| - \frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_N^2\mathbf{I})^{-1}\mathbf{y} \tag{3}$$

A commonly used kernel is the squared exponential (SE) defined as: $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\ell^2}\right)$, which depends on a length scale hyperparameter $\ell$. The optimization process usually involves optimization techniques such as gradient-based optimization or other optimization algorithms to maximize the log marginal likelihood with respect to the hyperparameters, including the noise variance $\sigma_N^2$ and the length scale $\ell$ for the SE kernel, to obtain the best-fitting hyperparameters for the given dataset.

In GPR predicting the mean $\mu_*$ and the variance $\sigma_*^2$ for a test point $\mathbf{x}_* \in \mathbb{R}^d$ involves the calculation of their respective mathematical expressions

$$\mu_* = \mathbf{k}_*^T(\mathbf{K} + \sigma_N^2\mathbf{I})^{-1}\mathbf{y} \tag{4}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_N^2\mathbf{I})^{-1}\mathbf{k}_* \tag{5}$$

where $\mathbf{K}$ is the covariance matrix of the training inputs $\mathbf{x}_i$, $\mathbf{k}_*$ is the vector of covariances between the test point $\mathbf{x}_*$ and training inputs $\mathbf{x}_i$, $k(\mathbf{x}_*, \mathbf{x}_*)$ is the variance of the test point $\mathbf{x}_*$, $\mathbf{I}$ is the identity matrix and $\mathbf{y}$ is the vector of outputs $y_i$.

To address the problem of GPR when dealing with large datasets, we introduce two clustering approaches based on selecting the most important points for GPR.

# 3 Reduced Gaussian process regression

Consider a large ($N \gg 1$) sample: $\mathcal{D} = \{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}_{i=1}^N$. The canonical GPR is expressed as

$$\begin{aligned} y_i &= f(\mathbf{x}_i) + \epsilon_i, \quad \text{with} \quad i = 1, \ldots, N \\ f(\mathbf{x}) &\sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \end{aligned} \tag{6}$$

Our challenge lies in managing the computational complexity of standard GPR while maintaining a good performance. Hence, we propose the use of two clustering approaches.

## 3.1 An efficient representation of observations

The objective is to identify a reduced subset of training observations that effectively capture the essential information within the data, facilitating both learning (3) and inference (4) tasks. Denote the number of clusters as $k$ with $k \ll N$. Initially, a clustering algorithm will group similar observations, resulting in clusters represented as $\mathcal{C} = \{C_j\}_{j=1}^{k}$ where each cluster $C_j$ contains a subset of training data points that are similar within the cluster $j$. Subsequently, each cluster $C_j$ will be represented by its center $\mathbf{c}_j \in \mathbb{R}^d$. These cluster centers will serve as a reduced set of observations. The reduced GPR model will then be learned on a reduced subset of observations, denoted as $\tilde{\mathcal{D}} = \{(\mathbf{c}_j, y_j)\}_{j=1}^{k}$, satisfying

$$\begin{aligned} y_j &= f(\mathbf{c}_j) + \epsilon_j, \quad \text{with} \quad j = 1, \ldots, k \\ f(\mathbf{c}) &\sim \mathcal{GP}(0, k(\mathbf{c}, \mathbf{c}')) \end{aligned} \tag{7}$$

**k-Means clustering.** The goal of the k-Means algorithm is to partition the data into $k$ clusters, such that each data point belongs to the cluster with the nearest centroid. Algorithm 1 iteratively assigns data points to the nearest centroid and updates the centroids until convergence, where convergence is typically defined as no change in the centroid assignments. Since the resulting centroids are usually not among the data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, we use the nearest neighbors (NN) classification method applied to the data in each cluster, and the nearest one will be assigned as the centroid $\mathbf{c}_j$.

**Algorithm 1** k-Means clustering.

1: **Input:** Data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and a number of clusters $k$
2: **Output:** centroids: $\mathbf{c}_1, \ldots, \mathbf{c}_k$
3: Initialize $k$ cluster centroids randomly: $\mathbf{c}_1^{(0)}, \ldots, \mathbf{c}_k^{(0)}$
4: **repeat**
5: For each $\mathbf{x}_i$ find $\text{cluster}(\mathbf{x}_i) = \arg\min_{1 \le l \le k} \|\mathbf{x}_i - \mathbf{c}_l^{(t)}\|^2$
6: Update the centroids by computing the mean of each cluster:

$$\mathbf{c}_l^{(t+1)} = \frac{1}{|\{\mathbf{x}_i \,|\, \text{cluster}(\mathbf{x}_i) = l\}|} \sum_{\{\mathbf{x}_i \,|\, \text{cluster}(\mathbf{x}_i) = l\}} \mathbf{x}_i$$

7: **until** convergence

**k-Medoids clustering.** The goal of the k-Medoids algorithm is to partition the data points into $k$ clusters, where each cluster is represented by a medoid. The medoid is defined as the data point within the cluster that minimizes the sum of distances to all other points in the same cluster. Algorithm 2 aims to minimize the total sum of distances between data points and their respective medoids.

## 3.2 The computational challenge

The canonical GPR in (6) exhibits a time complexity of $\mathcal{O}(N^3)$ for both learning and inference due to the need to invert the $N \times N$ covariance matrix $\mathbf{K}$, limiting its scalability for large

**Algorithm 2** k-Medoids clustering.

1: **Input:** Data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and a number of clusters $k$
2: **Output:** Medoids: $\mathbf{c}_1, \ldots, \mathbf{c}_k$
3: Initialize $k$ cluster medoids randomly: $\mathbf{c}_1^{(0)}, \ldots, \mathbf{c}_k^{(0)}$
4: **repeat**
5: For each $\mathbf{x}_i$ find cluster$(\mathbf{x}_i) = \arg\min_{1 \leq l \leq k} ||\mathbf{x}_i - \mathbf{c}_l^{(t)}||^2$
6: Update the medoids among $\mathbf{x}_i$ by selecting the point that minimizes the sum of distances within the cluster the medoid:

$$\mathbf{c}_l^{(t+1)} = \arg\min_{\mathbf{x}_i \in \mathbf{X}} \sum_{\mathbf{x} \in \text{cluster}(\mathbf{x}_i)} ||\mathbf{x} - \mathbf{x}_i||^2$$

7: **until** convergence

datasets when $N \gg 1$. Moreover, another limitation of GPR is the memory scaling $\mathcal{O}(N^2)$ in a direct implementation. In contrast, k-Means and k-Medoids clustering offer more efficient solutions. For k-Means clustering, the complexity is $\mathcal{O}(TkN)$, where $T$ represents the number of iterations. On the other hand, k-Medoids clustering yields a time complexity of $\mathcal{O}(k(N-k)^2)$. By employing k-Means or k-Medoids to select a reduced subset of representatives, the time complexity of the reduced GPR (RGPR) model in (7) can be substantially reduced to $\mathcal{O}(k^3)$ and the memory consumption to $\mathcal{O}(k^2)$ where $k \ll N$.

# 4 Experimental results

**Experimental protocol.** We employ a real dataset for binary classification comprising 1500 observations and a total of 200 features, illustrating the body temperature of dogs represented by their probability densities. Within this dataset, temporal measures of infected (501) and uninfected (999) dogs are recorded over a 24-hour period. Figure 1 shows the mean curve of each class with their confidence intervals of a level of 90%. We use 75% (1125 instances) for training and reserve 25% (375 instances) for test. For evaluation we consider three measures: Accuracy, F1-score, and AUC value.

The objective of this experiment is to evaluate the effectiveness of clustering-based methods, specifically k-Means and k-Medoids, in reducing the computational complexity of the standard GP classifier (GPC) while maintaining the predictive performance. For comparison, we deal with the PCA by extracting two principal directions for each observation (PCA GPC), which represent 98% of the initial information. In this investigation, we rely on the `sklearn` library (scikit-learn), which provides implementations of GPC, PCA, k-Means, and k-Medoids clustering algorithms.

**Results and discussion.** Table 1 shows the details of the training set for each case. In particular, the number of clusters for k-Means and k-Medoids is selected by employing the Elbow method for a fixed value range of [4,80]. The visualization in Figures 3, 2, 5, and 4 illustrates the reduced samples using two features selected in two directions through PCA.

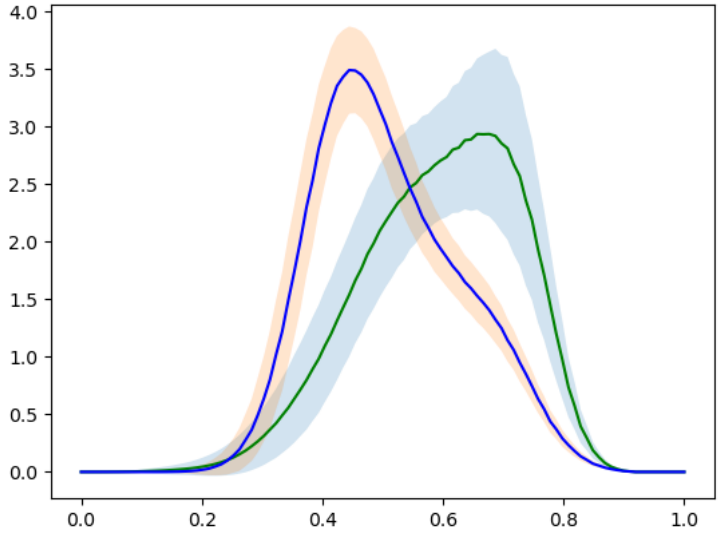



Figure 1: The mean curve of the first class (green) and the second class (blue) with their respective confidence intervals.

| **Method** | **Class 1** | **Class 2** | $(N, d)$ |
|---|---|---|---|
| GPC | 365 | 760 | (1125, 200) |
| PCA GPC | 365 | 760 | (1125, 2) |
| k-Means GPC | 59 | 66 | (125, 200) |
| k-Medoids GPC | 65 | 70 | (135, 200) |

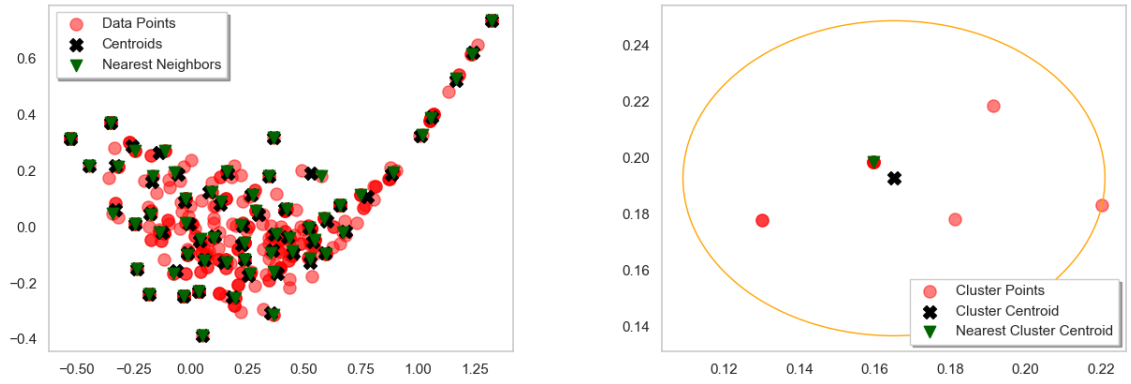
Table 1: Repartition and dimension of training set.

Specifically, it showcases centroids and their nearest points used as the reduced set for training in the case of k-Means, and medoids that will be utilized as the subset for model training in the case of k-Medoids. It is important to note that during the clustering process, we segregate the positive and negative classes, applying clustering to each class independently. Subsequently, we concatenate the reduced subsets of each class to obtain the final reduced training set.

**Comparison.** In our comparative study, we assess the performance of three models: The standard GPC, PCA GPC, and reduced GPC by both k-Means and k-Medoids. Figure 6 on the left displays the results over 100 random repetition, while Figure 6 on the right presents the mean values of the performance results. The mean execution times are summarized in Table 2.

# 5 Conclusion

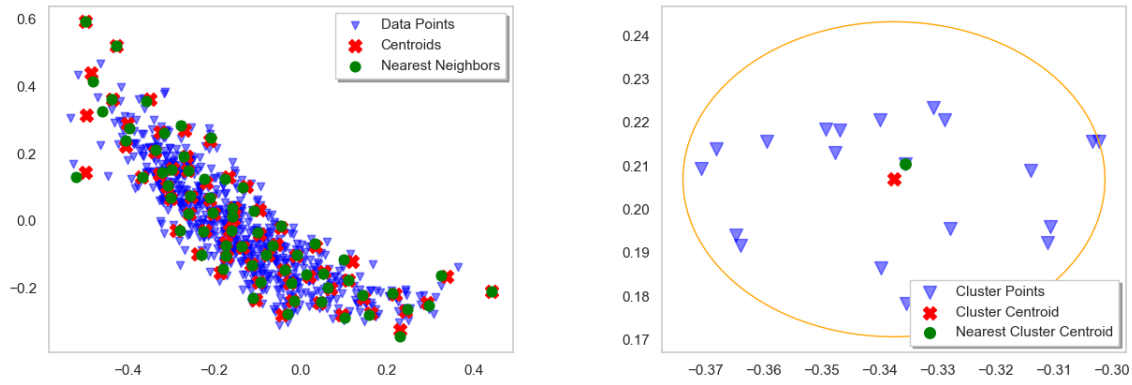
In this paper, we have proposed a new method for reducing the complexity and computation time of standard Gaussian processes for regression and classification through clustering




(a) Centroids of class 1

(b) One cluster of class 1

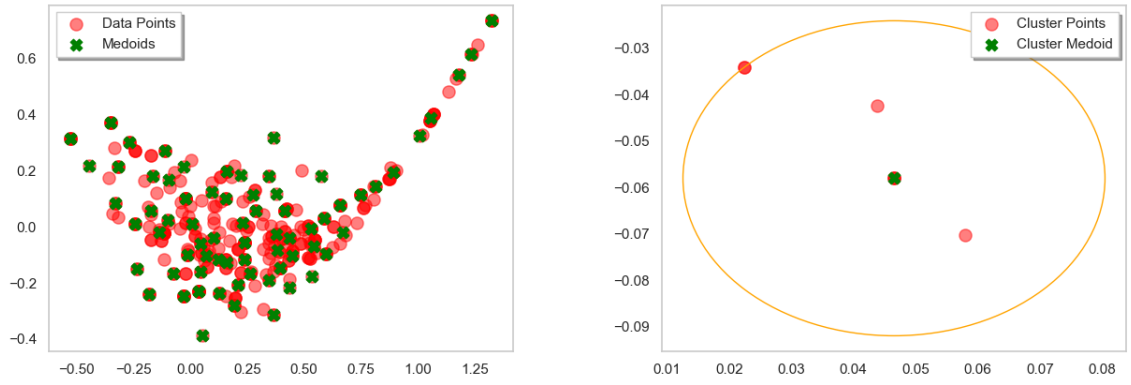
Figure 2: Centroids of class 1 using k-Means.




(a) Centroids of class 2

(b) One cluster of class 2

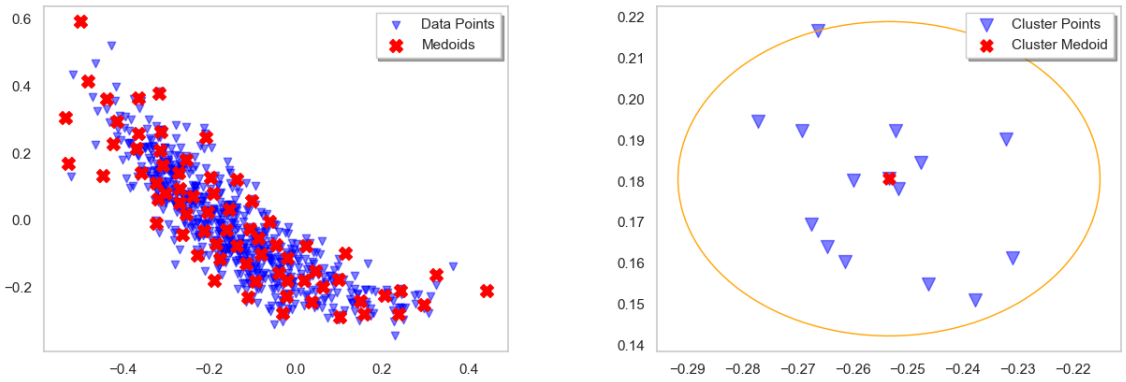
Figure 3: Centroids of class 2 using k-Means.




(a) Medoids of class 1

(b) One cluster of class 1

Figure 4: Medoids of class 1 using k-Medoids.




(a) Medoids of class 2

(b) One cluster of class 2

Figure 5: Medoids of class 2 using k-Medoids.

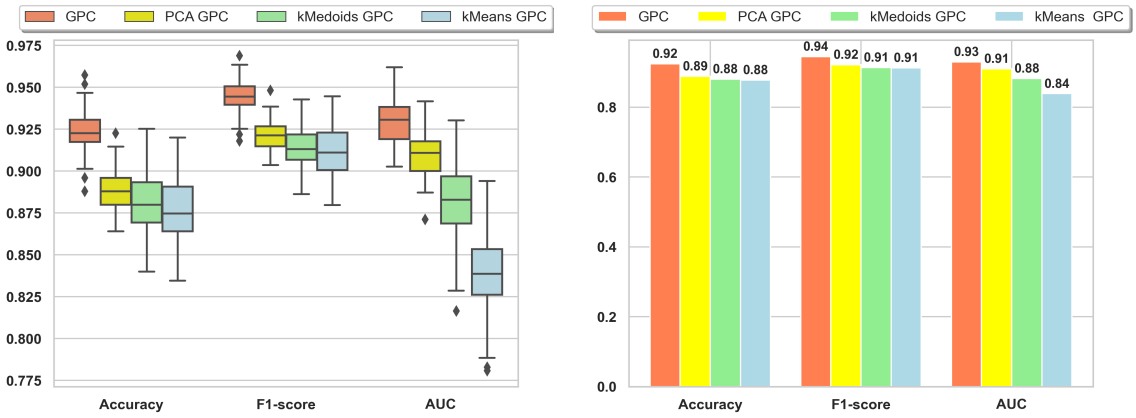



Figure 6: The left figure represents the box plots over 100 repetitions, while the right one illustrates the mean values for each metric.

| Method | Mean Execution Time (seconds) |
|---|---|
| GPC | 21.55 |
| PCA GPC | 12.48 |
| k-Means GPC | 1.17 |
| k-Medoids GPC | 0.27 |

Table 2: Mean of execution times for each method across 100 repetitions.

techniques. The experimental study confirmed that the proposed reduced Gaussian process has very promising advantages over the standard model in terms of complexity.

This link provides a code preview: Reduced-Run-Time-and-Memory-Complexity-of-GP.

# References


[1] A. Damianou, M. Titsias, and N. Lawrence. Variational Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2510–2518. Curran Associates, Inc., 2011.

[2] A. Fradi, C. Samir, and F. Bachoc. A scalable approximate Bayesian inference for high-dimensional Gaussian processes. *Communications in Statistics - Theory and Methods*, pages 1–68, 2020.

[3] A. Fradi, C. Samir, J. Braga, S. H. Joshi, and J-M. Loubes. Nonparametric Bayesian regression and classification on manifolds, with applications to 3D cochlear shapes. *IEEE Transactions on Image Processing*, 31:2598–2607, 2022.

[4] J. Fritz, W. Nowak, and I. Neuweiler. Application of FFT-based algorithms for large-scale universal kriging problems. *Mathematical Geosciences*, 51:199–221, 2009.

[5] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[6] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, pages 585–591. MIT Press, 2000.

# High-dimensional analysis of ridge regression for non-identically distributed data with a variance profile

Jérémie Bigot & Issa-Mbenard Dabo & Camille Male

*Institut de mathématiques de Bordeaux & CNRS (UMR 5251)*
*Université de Bordeaux, France*
*jeremie.bigot@math.u-bordeaux.fr; issa-mbenard.dabo@math.u-bordeaux.fr;*
*camille.male@math.u-bordeaux.fr*

**Résumé.** Les modèles de régression linéaire sont souvent étudiés dans le contexte de données indépendantes et identiquement distribuées. Nous proposons d'étudier ces modèles pour des données indépendantes mais non-identiquement distribuées. Dans ce but, nous supposons que l'ensemble des prédicteurs observés constitue une matrice aléatoire avec un profil de variance. En supposant un modèle à effets aléatoires, nous étudions le risque de prédiction de l'estimateur ridge pour la régression linéaire avec un tel profil de variance. Nous montrons un équivalent déterministe de ce risque, qui s'avère être un bon estimateur en grande dimension. Nous proposons également un équivalent déterministe du degré de liberté de l'estimateur ridge dans ce cadre. Nos travaux mettent aussi en évidence l'émergence du phénomène de double descente pour l'estimateur des moindres carrés de norme minimale lorsque le paramètre de régularisation ridge tend vers zéro. Les preuves de nos résultats s'appuient sur des outils issus de la théorie des matrices aléatoires en présence d'un profil de variance qui n'ont pas été considérés jusqu'à présent pour étudier les modèles de régression. Des expériences numériques sont fournies pour montrer l'exactitude de ces équivalents déterministes sur le calcul du risque de prédiction pour la régression ridge et des données non-identiquement distribuées.

**Mots-clés.** Régression ridge; Degrés de liberté; Double descente; Profil de variance; Hétéroscédasticité; Matrices aléatoires; Equivalents déterministes.

**Abstract.** Linear regression models are often studied in the context of independent and identically distributed data. We propose to investigate these models for independent but non-identically distributed data. To this end, we suppose that the set of observed predictors (or features) is a random matrix with a variance profile. Assuming a random effect model, we study the prediction risk of the ridge estimator for linear regression with such a variance profile. We provide a deterministic equivalent of this risk, which proves to be a good estimator in high-dimension. We also propose a deterministic equivalent of the degree of freedom of the ridge estimator in this setting. Our work also highlights the emergence of the double descent phenomenon for the minimum norm least-squares estimator when the ridge regularization parameter goes to zero. The proofs of our results are based on tools from random matrix theory in the presence of a variance profile that have not been considered so far to study regression models. Numerical experiments are provided to show the accuracy of the aforementioned deterministic equivalents on the computation of the prediction risk of ridge regression for non-identically distributed data.

**Keywords.** Linear ridge regression; Degrees of freedom; Double descent; Variance profile; Heteroscedasticity; Random Matrices; Deterministic equivalents.

# 1 Introduction

High dimensionality is a subject of interest in the field of statistics, especially in regression problems, driven by the advent of big data. This context gives rise to unexpected phenomena and contradictions with established statistical heuristics when the dimension $p$ of the predictors is fixed and the number $n$ of observations tends to infinity. These phenomena appear particularly in the context of linear regression. Indeed, as the sample size and dimension of acquired data increase, the study of this model is different from the classical framework. Indeed, in the asymptotic regime where $\min(n,p) \to +\infty$ and $\frac{p}{n} \to c > 0$, one can notably mention the obsolescence of certain estimators, the occurrence of double descent, or examples of overfitting. In this asymptotic setting, using tools from random matrix theory (RMT), many authors have therefore focused on the consequences of high-dimensionality on linear regression, see e.g. [DW18, Bac23, HMRT22, LC18]. In this paper, we focus on the linear regression model

$$Y_n = X_n\theta_* + \varepsilon_n, \tag{1.1}$$

where $X_n$ is $n \times p$ matrix of random predictors, $\varepsilon_n \in \mathbb{R}^n$ is a noise vector independent of $X_n$ with $\mathbb{E}[\varepsilon_n] = 0$ and $\mathbb{E}[\varepsilon_n \varepsilon_n^T] = \sigma^2 I_n$, $\theta_* \in \mathbb{R}^p$ is a vector of unknown parameters, and $Y_n \in \mathbb{R}$ is the vector of observed responses.

Classically, the predictors are assumed to be independent and identically distributed (iid) data, meaning that the rows of the matrix $X_n$ are independent vectors sampled from the same probability distribution. In this paper, we propose to depart from this assumption by considering the setting where the rows of $X_n$ are independent but non-identically distributed. To this end, we suppose that $X_n$ is expressed in the following form

$$X_n = \Upsilon_n \circ Z_n,$$

where $\circ$ denotes the Hadamard (entry-wise) product between two matrices, $Z_n = (Z_{ij})$ has iid centered entries with variance one, and $\Upsilon_n = (\gamma_{ij})$ is a deterministic matrix. The matrix $(\gamma_{ij}^2) \in \mathbb{R}^{n\times p}$ governs the variance of the entries of $X_n$, and it is called a variance profile. The motivation for studying linear regression using such a variance profile is to consider the setting where one has $n$ independent pairs of observations $(Y_i, X_i)_{1\le i\le n}$ (with $X_i = (X_{ij})_{1\le j\le p}$) that are not necessarily identically distributed. Note that in the standard setting of iid data, one that

$$\gamma_{ij} = \gamma_j \quad \text{for all} \quad 1 \le i \le n, \text{ and } 1 \le j \le p.$$

The main goal of this paper is then to understand how assuming such a variance profile for $X_n$ influences the statistical properties of ridge regression in the linear model (1.1) when compared to the standard assumption of iid observations. In this setting, our approach also allows to analyze the performances of the minimum norm least-squares estimator when the ridge regularization parameter goes to zero.

We consider the possibly high-dimensional context (with $p \geq n$) when the least squares estimator is not uniquely defined. Thus, we focus our analysis on the ridge regression estimator that is the minimizer of the following loss function

$$\hat{\theta}_\lambda = \underset{\theta \in \mathbb{R}^p}{\arg\min} \frac{1}{n}\|Y_n - X_n\theta\|^2 + \lambda\|\theta\|^2,$$

for some regularization parameter $\lambda > 0$. Regardless of the ratio between $n$ and $p$, this estimator has the following explicit expression

$$\hat{\theta}_\lambda = (X_n^T X_n + n\lambda I_p)^{-1} X_n^T Y_n = X_n^T (X_n X_n^T + n\lambda I_n)^{-1} Y_n.$$

Our analysis also includes the study of the minimum norm least-squares estimator defined as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\arg\min} \left\{ \|\theta\| \ : \ \theta \text{ minimizes } \frac{1}{n}\|Y_n - X_n\theta\|^2 \right\}$$

to which the ridge regression estimator converges when $\lambda$ tends to zero. Note that the minimum norm least-squares estimator is also known to be the limit of gradient descent when started at zero [GLSS18].

To study the statistical performances of the ridge regression estimator, we analyse its predictive risk defined as

$$\hat{r}_\lambda(X_n) = \mathbb{E}[(\tilde{Y} - \tilde{X}^T\hat{\theta}_\lambda)^2 | X_n], \tag{1.2}$$

where $(\tilde{Y}, \tilde{X}) \in \mathbb{R} \times \mathbb{R}^p$ is independent from $(Y_n, X_n)$ and satisfies

$$\tilde{Y} = \tilde{X}^T\theta_* + \tilde{\varepsilon}, \text{ with } \mathbb{E}[\tilde{\varepsilon}] = 0, \ \mathbb{E}[\tilde{\varepsilon}^2] = \sigma^2.$$

In the above formula, $\tilde{X} = \tilde{\Gamma}_p^{1/2}\tilde{Z}$ with $\tilde{Z} \in \mathbb{R}^p$ a random vector with iid centered entries and variance one and $\tilde{\Gamma}_p = \mathbb{E}[\tilde{X}\tilde{X}^T] = \mathrm{diag}(\tilde{\gamma}_1^2, \ldots, \tilde{\gamma}_p^2)$ denotes the variance profile of $\tilde{X}$. Note that the risk $\hat{r}_\lambda(X_n)$ is conditioned on the predictors $X_n$, and it is thus a random variable.

Following [DW18], we focus on a random-effect hypothesis that assumes that the components of the vector $\theta_*$ are drawn independently at random. As argued in [DW18], this corresponds to an average case analysis over a set of dense regression coefficients as opposed to the "sparsity hypothesis" [HTW15] or the "manifold hypothesis" [LHT23] that are other popular assumptions in high-dimensional linear regression. More precisely, the following random coefficients assumption is made throughout the paper.

**Assumption 1.1** *The vector $\theta_*$ of regression coefficients is random, independent from $X_n$, $\tilde{X}$, $\varepsilon_n$ and $\tilde{\varepsilon}$, with $\mathbb{E}[\theta_*] = 0$ and*

$$\mathbb{E}[\theta_*\theta_*^T] = \frac{\alpha^2}{p} I_p.$$

The above coefficient $\alpha > 0$ represents the average amount of signal strength in model (1.1).

## 1.1 Main contributions

Recall that the prediction of $Y_n$ by ridge regression is

$$\hat{Y}_\lambda = X_n \hat{\theta}_\lambda = A_\lambda Y_n, \quad \text{where} \quad A_\lambda = X_n (X_n^T X_n + n\lambda I_p)^{-1} X_n^T.$$

Then, the so-called degrees of freedom (DOF) of the estimator $\hat{\theta}_\lambda$, that is defined as

$$\widehat{df}_1(\lambda) = \mathrm{Tr}[A_\lambda] = \mathrm{Tr}[\widehat{\Sigma}_n (\widehat{\Sigma}_n + \lambda I_p)^{-1}], \text{ where } \widehat{\Sigma}_n = \frac{1}{n} X_n^T X_n,$$

represents the so-called effective dimension of the linear estimator $\hat{Y}_\lambda$. Inspired by recent results from [Bac23] in the setting of iid data, a first contribution of this work is to prove the following deterministic equivalence of the DOF

$$\widehat{df}_1(\lambda) \sim df_1(\lambda), \quad \text{where} \quad df_1(\lambda) = \mathrm{Tr}[\Gamma_n (\Gamma_n + \kappa(\lambda))^{-1}], \tag{1.3}$$

with

$$\Gamma_n = \mathbb{E}[\widehat{\Sigma}_n] = \frac{1}{n} \mathrm{diag}\left( \sum_{i=1}^n \gamma_{i1}^2, \ldots, \sum_{i=1}^n \gamma_{ip}^2 \right),$$

and $\kappa(\lambda)$ is a diagonal matrix that depends upon the regularization parameter $\lambda$ and the variance profile matrix.

Hence, the equivalence relation (1.3), to be understood as

$$\lim_{n\to\infty,\, p/n\to c} |\widehat{df}_1(\lambda) - df_1(\lambda)| = 0, \text{ almost surely},$$

indicates that the DOF of the ridge regression estimator for the empirical covariance matrix $\widehat{\Sigma}_n$ corresponds to the DOF computed with the population covariance matrix $\Gamma_n$ and another additive regularization structure than $\lambda I_p$ that is given by the diagonal matrix $\kappa(\lambda)$ whose explicit expression is given in Section 3.

Then, the second and main contribution of the paper is to derive a deterministic equivalent of the predictive risk $\hat{r}_\lambda(X_n)$ in the case where the number of samples $n$ and the dimension $p$ tend to infinity at a proportional rate. This deterministic equivalent allows us to understand the influence of the ratio $c$ on the predictive risk and to also analyze the effect of the signal strength $\alpha$. We also study the convergence of the predictive risk as $\lambda$ tends 0 to analyze the statistical properties of the minimum norm least square estimator. In this setting, it appears a phenomenon arising from the curse of dimensionality that is commonly known as double descent for iid data. This phenomenon contradicts the consensus heuristic that, when a model becomes over-parameterized, then the predictive risk increases due to overfitting of the training data and the model is no longer capable of generalizing. This double descent has been thoroughly studied in the case of high-dimensional linear regression using tool from RMT, see e.g. [HMRT22, Bac23, BHX20a] and references therein. In this paper, we show that it also occurs for non iid data with a variance profile.

As a third contribution, using synthetic data and illustrative examples of variance profile, we conduct various numerical experiments to verify the accuracy of our deterministic equivalents of the DOF and the predictive risk using finite samples. We also investigate the similarities and differences that exist between the standard setting of iid data and the one of non-identically distributed data with a variance profile.

# 2 Related works

## 2.1 High-dimensional linear regression from the random matrix perspective

In the setting where the sample size is comparable to the dimensionality of the observations, recent advances in random matrix theory (RMT) have been successfully applied to various inference problems in high-dimensional multivariate statistics, see e.g. [NPW21] for a recent overview. Many works have considered the high-dimensional analysis of the linear model using tools from RMT for iid data with a general covariance structure $\Sigma \in \mathbb{R}^{p\times p}$ (assumed to be a positive semi-definite matrix) that is for

$$X_n = Z_n \Sigma^{1/2}, \text{ for an } n \times p \text{ matrix } Z_n \text{ with iid centered entries having variance one.}$$

In particular, for such data, the study of the minimum norm least-squares estimator and the double descent behavior of the predictive risk has been considered in [HMRT22, Bac23, BHX20b, RMR21]. The analysis of the predictive risk of ridge regression using iid data with a general covariance structure has been studied in [DW18, Bac23], while previous works on the statistical analysis of ridge regression from the RMT perspective include [EK18, Dic16] and [CD11, TV04] for applications in wireless communication.

## 2.2 Linear regression for independent but non-identically distributed data

While the statistical analysis of linear regression for iid data with a general covariance structure is very well understood, the literature on the study of the linear model for non-identically distributed predictors appears to be rather scarce. A first analysis of maximum likelihood estimation in standard models (including linear regression) for independent but non-identically distributed data dates back to [Ber82]. More recent works [BBK+19, KBB+20] on statistical inference in linear regression in the so-called model-free framework allow to consider the setting on non-identically distributed predictors. However, to the best of our knowledge, the high-dimensional analysis of the linear model using non-identically distributed data has not considered so far.

In this paper, we build upon results from [HLN07] to construct deterministic equivalents of the Stieltjes transforms

$$g_{\hat{\mu}_n}(z) = \frac{1}{p}\operatorname{Tr}[(\widehat{\Sigma}_n - zI_p)^{-1}] \text{ and } g_{\tilde{\mu}_n}(z) = \frac{1}{n}\operatorname{Tr}[(\widetilde{\Sigma}_n - zI_n)^{-1}] \quad \text{for} \quad z \in \mathbb{C}\setminus\mathbb{R}^+,$$

of the empirical eigenvalue distribution $\hat{\mu}_n$ of $\widehat{\Sigma}_n$, and the empirical eigenvalue distribution $\tilde{\mu}_n$ of $\widetilde{\Sigma}_n = \frac{1}{n}X_nX_n^T$ respectively, when $X_n = \Upsilon_n \circ Z_n$ has a variance profile $(\gamma_{ij}^2)$.

**Proposition 2.1** *The Stieltjes transforms $g_{\hat{\mu}_n}(z)$ and $g_{\tilde{\mu}_n}(z)$ satisfy*

$$\lim_{n\to\infty,\, p/n\to c}\left(g_{\hat{\mu}_n}(z) - \frac{1}{p}\operatorname{Tr}[T(z)]\right) = 0, \quad \textit{for all} \quad z \in \mathbb{C}\setminus\mathbb{R}^+,$$

$$\lim_{n\to\infty,\, p/n\to c} \left( g_{\tilde{\mu}_n}(z) - \frac{1}{n}\operatorname{Tr}[\tilde{T}(z)] \right) = 0, \quad \textit{for all} \quad z \in \mathbb{C} \setminus \mathbb{R}^+,$$

*where*

$$T(z) = \operatorname{diag}(T_1(z), \ldots, T_p(z)) \quad \textit{and} \quad \tilde{T}(z) = \operatorname{diag}(\tilde{T}_1(z), \ldots, \tilde{T}_n(z)),$$

*are diagonal matrices of size $p \times p$ and $n \times n$ respectively, whose diagonal elements are the unique solutions of the deterministic system of $p + n$ equations*

$$T_j(z) = \frac{-1}{z\left(1 + (1/n)\operatorname{Tr}[\tilde{D}_j \tilde{T}(z)]\right)} \quad \textit{for} \quad 1 \le j \le p, \tag{2.1}$$

$$\tilde{T}_i(z) = \frac{-1}{z\left(1 + (1/n)\operatorname{Tr}[D_i T(z)]\right)} \quad \textit{for} \quad 1 \le i \le n, \tag{2.2}$$

*where*

$$\tilde{D}_j = \operatorname{diag}(\gamma_{1j}^2, \ldots, \gamma_{nj}^2) \quad \textit{and} \quad D_i = \operatorname{diag}(\gamma_{i1}^2, \ldots, \gamma_{ip}^2).$$

*Moreover, $\frac{1}{p}\operatorname{Tr}[T(z)]$ and $\frac{1}{n}\operatorname{Tr}[\tilde{T}(z)]$ are the Stieltjes transforms of probability measures denoted as $\nu_n$ and $\tilde{\nu}_n$ that are the deterministic equivalents of $\hat{\mu}_n$ and $\tilde{\mu}_n$ respectively.*

# 3 Main results

In this section, we derive deterministic equivalents for the DOF and the predictive risk of ridge regression. We also obtain a deterministic equivalent of the predictive risk of minimum norm least square estimation when the ridge regularization parameter tends to zero. We compare these results to those that are already known in the standard setting of iid data, and we highlight the emergence of the double descent phenomenon for non-iid data.

**Assumption 3.1** *There exists $\delta > 0$ such that, $\mathbb{E}[|Z_{ij}|^{4+\delta}] < +\infty$.*

**Assumption 3.2** *There exists $\gamma_{\max} > 0$ such that, $\sup_{n\ge 1} \max_{\substack{1\le i\le n \\ 1\le j\le n}} |\gamma_{ij}| < \gamma_{\max}$.*

**Assumption 3.3** *There exists $\gamma_{\min} > 0$ such that, $\forall n \ge 1$, $\min_{\substack{1\le i\le n \\ 1\le j\le n}} |\gamma_{ij}| \ge \gamma_{\min}$.*

## 3.1 Predictive risk

**Theorem 3.1** *Consider the linear model (1.1). Then, under Assumptions (3.1) and (3.2), a deterministic equivalent of the predictive risk is*

$$r_\lambda(\Upsilon_n) = \sigma^2 + \frac{\sigma^2}{n}\operatorname{Tr}[\tilde{\Gamma}_p T(-\lambda)] + \lambda\left(\frac{\lambda\alpha^2}{p} - \frac{\sigma^2}{n}\right)\operatorname{Tr}[\tilde{\Gamma}_p T'(-\lambda)],$$

*in the sense that it satisfies*

$$\lim_{n\to\infty,\, p/n\to c} \hat{r}_\lambda(X_n) - r_\lambda(\Upsilon_n) = 0 \quad \text{a.s.}$$

*Moreover, for* $\lambda > 0$ *and* $\lambda_* = \frac{\sigma^2 p}{\alpha^2 n}$, $r_{\lambda_*}(\Upsilon_n) \leq r_\lambda(\Upsilon_n)$ a.s.

Note that our deterministic equivalent $r_\lambda(\Upsilon_n)$ of the predictive risk allows to derive an optimal choice $\lambda_*$ for the regularization parameter that corresponds to the one obtained for iid data in [DW18], and that is independent of the variance profile. Theorem 3.1 also allows us to understand the behavior of $r_\lambda(\Upsilon_n)$ when $\lambda \to 0$ and $\lambda \to +\infty$ through the following corollary.

**Corollary 3.1** *Under Assumptions (3.1), (3.2) and (3.3), the limit of the deterministic equivalent* $r_\lambda(\Upsilon_n)$ *for* $\lambda \to +\infty$ *and* $\lambda \to 0$ *are as follows*

$$\lim_{\lambda\to+\infty} r_\lambda(\Upsilon_n) = \frac{\alpha^2}{p}\mathrm{Tr}[\tilde{\Gamma}_p] + \sigma^2.$$

*If* $\frac{p}{n} < 1$, *then*

$$\lim_{\lambda\to 0} r_\lambda(\Upsilon_n) = \frac{\sigma^2}{n}\mathrm{Tr}[\tilde{\Gamma}_p T(0)] + \sigma^2.$$

*If* $\frac{p}{n} > 1$, *then*

$$\lim_{\lambda\to 0} r_\lambda(\Upsilon_n) = \frac{\alpha^2}{p}\mathrm{Tr}[\tilde{\Gamma}_p \kappa(0)(\tilde{\Gamma}_p + \kappa(0))^{-1}] + \frac{\sigma^2}{n}\mathrm{Tr}[\kappa'(0)\tilde{\Gamma}_p \Delta_n (\tilde{\Gamma}_p + \kappa(0))^{-2}] + \sigma^2,$$

*where for* $\lambda \in \mathbb{R}^+$

$$\kappa(\lambda) = \operatorname*{diag}_{1\leq j\leq p}\left(\frac{\mathrm{Tr}[\tilde{D}_j]}{\mathrm{Tr}[\tilde{D}_j \tilde{T}(-\lambda)]}\right) \quad \text{and} \quad \Delta_n = \frac{1}{n}\operatorname{diag}(\mathrm{Tr}[\tilde{D}_1], \ldots, \mathrm{Tr}[\tilde{D}_p])$$

# 4 Numerical experiments

In this section, we illustrate our results with numerical experiments. Figure 1 compares the predictive risk and its deterministic equivalent. The red curve represent the deterministic equivalent $r_\lambda(\Upsilon_n)$ whereas the black one represents the actual predictive risk $\hat{r}_\lambda(\Upsilon_n)$. On the other had, the dotted line depicts $(\tilde{Y} - \tilde{X}^T\hat{\theta}_\lambda)^2$, which is a realisation of the predictive risk. Figure 1 confirms that $r_\lambda(\Upsilon_n)$ is a good estimator of $\hat{r}_\lambda(\Upsilon_n)$ in high dimension as the black and the red curves coincide. The dotted line stays around the two other curves as they estimate $\hat{r}_\lambda(\Upsilon_n)$ which is the expectation of $(\tilde{Y} - \tilde{X}^T\hat{\theta}_\lambda)^2$.

Figure 2 represents the predictive risk with for different values of the ratio $n/p$, the solid lines depicts the case with $\lambda = 0$ and the dashed lines depicts the case with $\lambda = \lambda_* = \frac{\sigma^2 p}{\alpha^2 n}$. The double descent phenomenon is illustrated by the solid lines since the curves are increasing for $p/n < 1$ and decreasing for $p/n > 1$. This is related to Corollary 3.1 which states that the expression of $\lim_{\lambda\to 0} r_\lambda(\Upsilon_n)$ depends upon the value of the ratio $p/n$ with respect to one. The performances of $\hat{\theta}_0$ and $\hat{\theta}_*$ seem similar when $p/n$ is large whereas the performances obtained with $\hat{\theta}_*$ are much better than with those using $\hat{\theta}_0$.

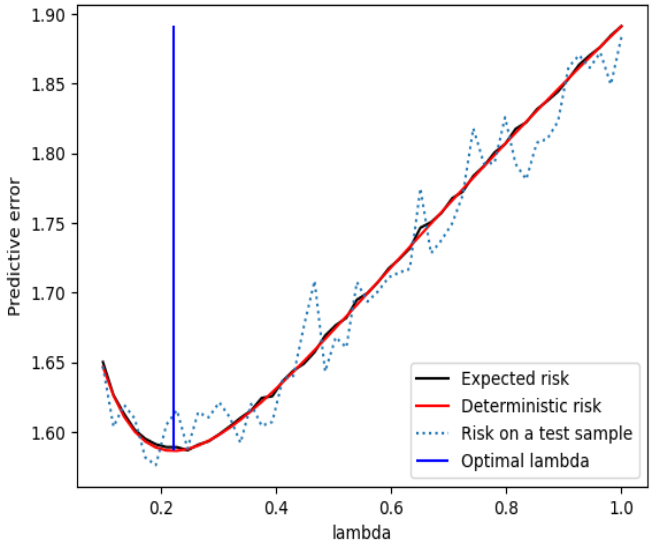



Figure 1: Comparison of $r_\lambda(\Upsilon_n)$ and $\hat{r}_\lambda(\Upsilon_n)$ with a doubly-stochastic variance profile (rows and columns sum up to the same constant value), $\alpha = 1.5$, $\sigma = 1$, $n = 800$ and $p = 500$,

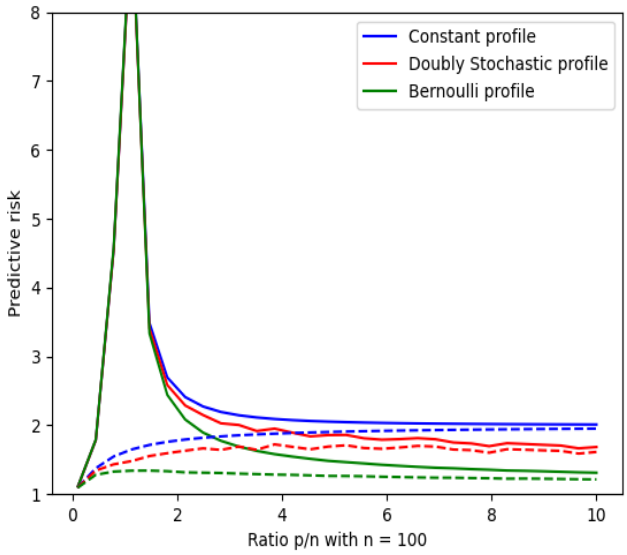



Figure 2: Double descent phenomenon for different variance profile with $\alpha = \sigma = 1$, $n = 100$ and $p$ varying from 10 to 500. The Bernoulli variance profile corresponds to variance $\gamma_{ij}$ randomly sampled from a Bernoulli distribution. The full lines correspond to $\lambda = 0$ and the dashed lines correspond to $\lambda = \lambda_*$.

# 5 Conclusion

In this paper, we have derived a deterministic equivalent of the DOF and the predictive risk of ridge (less) regression in a high-dimensional framework with a variance profile to handle the setting of non-iid data. The numerical experiments that we have conducted confirm that this deterministic equivalent accurately estimates the predictive risk in high-dimension. Our results also allow to understand how assuming such a variance profile for the data influences the statistical properties of ridge regression when compared to the standard assumption of iid observations. We hope that our approach on the use of variance profiles may lead to further research works on the statistical analysis of other estimators than ridge regression in more complex models with non-iid data.

# References


[Bac23] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023.

[BBK+19] Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as Approximations II: A Model-Free Theory of Parametric Regression. *Statistical Science*, 34(4):545 – 565, 2019.

[Ber82] Rudolf Beran. Robust Estimation in Models for Independent Non-Identically Distributed Data. *The Annals of Statistics*, 10(2):415 – 428, 1982.

[BHX20a] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[BHX20b] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[CD11] Romain Couillet and Mérouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[Dic16] Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1 – 37, 2016.

[DW18] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.

[EK18] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95–175, 02 2018.

[GLSS18] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas

Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR, 10–15 Jul 2018.

[HLN07] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875 – 930, 2007.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[KBB+20] Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, Junhui Cai, Edward I. George, and Linda H. Zhao. Valid post-selection inference in model-free linear regression. *The Annals of Statistics*, 48(5):2953 – 2981, 2020.

[LC18] Zhenyu Liao and Romain Couillet. The dynamics of learning: A random matrix approach. In *International Conference on Machine Learning*, pages 3072–3081. PMLR, 2018.

[LHT23] Liangchen Liu, Juncai He, and Richard Tsai. Linear regression on manifold structured data: the impact of extrinsic geometry on solutions, 2023.

[NPW21] Jamshid Namdari, Debashis Paul, and Lili Wang. High-dimensional linear models: A random matrix perspective. *Sankhya A: The Indian Journal of Statistics*, 83(2):645–695, 2021.

[RMR21] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge(less) regression under general source condition. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 3889–3897. PMLR, 2021.

[TV04] Antonia Maria Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Found. Trends Commun. Inf. Theory*, 1(1), 2004.

# Modèles de régression ordinal cumulatif à covariables temporelles

Simón Weinberger [1] & Jairo Cugliari [2] & Aurélie Le Cain [3]

[1] *Laboratoire ERIC, EssilorLuxottica, France, weinbes@essilor.fr*
[2] *Laboratoire ERIC, France, Jairo.Cugliari@univ-lyon2.fr*
[3] *EssilorLuxottica, France, lecaina@essilor.fr*

**Résumé.** Dans ce document, on s'intéresse à la modélisation d'une réponse ordinale en utilisant des extensions du modèle ordinal cumulatif multivarié, présenté dans le livre de Agresti, A. (2010). On porte notre attention sur l'utilisation de covariables qui dépendent du temps. Pour cela, on explore deux représentations temporelles de nature différente : la représentation fonctionnelle présentée dans le monographe de Ramsay, J. O. et Silverman, B. W. (2005) et la signature du signal, presenté dans les notes Chevyrev, I. et Kormilitzin, A. (2016) ou encore la thèse de Fermanian, A. (2021). La principale contribution de cet article est l'utilisation d'une pénalité fonctionnelle dans le modèle ordinal à covariables fonctionnelles présenté par Jacques, J. et Samardzić, S. (2022). On montre la pertinence des méthodes proposées en analysant deux jeux de données publiques.

**Mots-clés.** Données fonctionnelles, signature, régularisation.

**Abstract.** We focus on modeling an ordinal response using extensions of the multivariate cumulative ordinal model, as presented in Agresti, A.'s (2010) book. Our attention is drawn to the use of covariates that depend on time. To this end, we explore two temporally based representations of different natures : the functional representation outlined in Ramsay, J. O., and Silverman, B. W.'s (2005) monograph, and the signal signature, introduced in Chevyrev, I., and Kormilitzin, A.'s (2016) notes, as well as in Fermanian, A.'s (2021) thesis. The primary contribution of this article lies in the application of a functional penalty in the ordinal model with functional covariates, as presented by Jacques, J., and Samardzić, S. (2022). We demonstrate the relevance of the proposed methods by analyzing two public datasets.

**Keywords.** Functional data, signatures, regularization.

Ce document est structuré de la façon suivante : dans la section 1 on définit les différentes représentations temporelles, on formule des modèles de régression appropriés et finalement, on s'intéresse à l'estimation des paramètres du modèle ordinal à covariables fonctionnelles. Dans la section 2, on applique les méthodes présentées à deux jeux de données publiques.

# 1 Méthode

## 1.1 Représentation de données temporelles

On considère une réponse discrète $Y$ qui peut prendre $K \in \mathbb{N}$ valeurs distinctes. On suppose qu'il existe une relation d'ordre total dans cet ensemble de valeurs, on peut alors supposer, sans perdre de généralité, que $Y$ prend des valeurs dans $[\![1, K]\!]$. On dit alors que $Y$ est une réponse ordinale.

On s'intéresse alors à modéliser la variable $Y$ en utilisant une covariable $Z(.)$, qui est un processus aléatoire dépendant du temps. Par le théorème de décomposition de Wold, il existe une fonction déterministe $X(.)$ et un processus aléatoire centré, $\epsilon(.)$, tel que :

$$Z(t) = X(t) + \epsilon(t) \quad ; \quad t \in [0, T].$$

Dans cet article, on explore l'utilisation de $X(.)$ ou de $\epsilon(.)$ comme variables explicatives pour la réponse $Y$. Ces deux variables dépendent du temps, mais sont de nature différente : la fonction $X(.)$ décrit le niveau moyen de $Z(.)$ et le processus $\epsilon(.)$ décrit l'écart entre ce niveau moyen et les valeurs de $Z(.)$. Deux approches, déjà existantes dans la littérature, sont d'intérêt : dans l'article de Jacques, J. et Samardzić, S. (2022), un modèle ordinal à covariables fonctionnelles est décrit et dans les notes de Chevyrev, I. et Kormilitzin, A. (2016), une méthode pour classifier des processus ARMA est proposée.

Dans le domaine des données fonctionnelles, il est usuel de représenter la fonction $X(.)$ en utilisant un développement fini dans une base fonctionnelle. Concrètement, on considère que la fonction $X(.)$ peut s'écrire de la façon suivante :

$$X(t) = \sum_{k=1}^{K} a_k \psi_k(t) = \boldsymbol{a^T} \cdot \boldsymbol{\psi}(t) \quad ; \quad t \in [0, T],$$

où $(\psi_k)_{i=1}^K$ est une base fonctionnelle connue (base de Fourrier, splines, ondelettes ...), $\boldsymbol{a} = (a_1, \ldots, a_K)^T$ est un vecteur de coefficients et $\boldsymbol{\psi}(t) = (\psi_1(t), \ldots, \psi_K(t))$ est un vecteur de fonctions.

Comme on illustre dans la section 2, dans certaines situations, il y a de l'information utile pour la prédiction de la réponse $Y$ dans le processus $\epsilon(.)$. La représentation fonctionnelle ne prend pas en compte ce processus, il est souvent considéré comme du bruit, or, il se pourrait que ce bruit soit informatif.

On présente succinctement une méthode, proposée dans Chevyrev, I. et Kormilitzin, A. (2016), pour extraire de l'information de $\epsilon(.)$. Cette technique consiste à transformer le processus univarié $\epsilon(.)$ dans un processus bivarié $\tilde{\epsilon(.)}$ et ensuite, on utilise la "signature" de ce processus, à valeurs dans $\mathbb{R}^2$, pour extraire des caractéristiques utiles pour la régression ordinale. Définissons la signature d'un processus aléatoire, $\tilde{\epsilon}(.)$, à valeurs dans $\mathbb{R}^d$ : $\tilde{e}(.) = (\tilde{e}_1(.), \tilde{e}_2(.), \ldots, \tilde{e}_d(.))^T$.

Soit $\omega_k$ un multi-index, de taille $k \in \mathbb{N}$, $\omega_k = (i_1, i_2, \ldots, i_k) \in \{1, \ldots, d\}^k$, l'intégrale itérée

de $\tilde{\epsilon}$ dans l'intervalle $[a, b]$ par rapport à $\omega_k$, est la quantité suivante :

$$S^{\omega_k}_{[a,b]}(\tilde{\epsilon}) = \int\limits_{a<s_1<\ldots<s_k<b} \ldots \int d\tilde{e}_{i_1} \ldots d\tilde{e}_{i_k}.$$

La signature de $\tilde{\epsilon}$ de niveau $k \in \mathbb{N}$, est l'ensemble des intégrales itérées par rapport aux multi-index de longueur plus petite ou égale à $k$. On note cet ensemble $S^{(k)}_{[a,b]}$. La signature de $\tilde{\epsilon}$ est l'ensemble des intégrales itérées. On note cet ensemble-là par $S_{[a,b]}(\tilde{\epsilon})$. Dans la thèse de Fermanian, A. (2021), il a été proposé d'utiliser la signature (de niveau $k$ fixe) pour extraire des caractéristiques dans des tâches d'apprentissage automatique où la covariable est un de processus multivarié. Dans le cas univarié, il est conseillé de transformer le processus dans un processus multivarié en appliquant différentes transformations.

Par exemple, pour des données univariées $\mathcal{M} = \{x_1, \ldots x_n\}$, on considère les transformations suivantes : la somme cumulée, notée $CS(.)$, le lead-lag, $LeadLag(.)$ et le point de base $BP(.)$, sont les transformations suivantes des données $\mathcal{M}$ :

$$\begin{aligned}
CS(\mathcal{M}) &= \{x_1, \sum_{i=1}^{2} x_i, \sum_{i=1}^{3} x_i, \sum_{i=1}^{3} x_i, \ldots, \sum_{i=1}^{n} x_i\}, \\
LeadLag(\mathcal{M}) &= \left\{ \begin{pmatrix} x_1 \\ x_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ x_2 \end{pmatrix}, \begin{pmatrix} x_3 \\ x_2 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ x_{n-1} \end{pmatrix}, \begin{pmatrix} x_n \\ x_n \end{pmatrix} \right\}, \\
BP(\mathcal{M}) &= \{0, x_1, x_2, \ldots, x_n\}.
\end{aligned}$$

Dans la pratique, on n'observe pas la fonction $Z(.)$, mais des données de la forme $\{(t_i, Z(t_i))\}_{i=1}^{T}$. Pour estimer la fonction $X(.)$, on fixe une base fonctionnelle $(\psi_k(.))_{k=1}^{K}$ donnée, et il faut déterminer les coefficients, $(a_k)_{k=1}^{K}$. Une approche possible, qui est décrite par Ramsay, J. O. et Silverman, B. W. (2005), est de choisir les coefficients qui minimisent la quantité suivante :

$$\sum_{i=1}^{T} \left( Z(t_i) - \hat{X}(t_i) \right)^2 + \lambda_X \int_{[0,T]} \left( \hat{X}^{(k)}(s) \right)^2 ds,$$

où $\hat{X}(t) = \sum_{k=1}^{K} a_k \psi_k(t)$ est l'estimation de la fonction $X(.)$ et $\lambda_X > 0$ est un réel positif connu qui contrôle l'irrégularité de $\hat{X}(t)$.

Une fois que l'on a cette représentation fonctionnelle, on doit estimer le processus $\epsilon(.)$, pour cela, on propose d'utiliser les observations initiales $\{(t_i, Z(t_i))\}_{i=1}^{T}$ et $\{(t_i, \hat{X}(t_i))\}_{i=1}^{T}$. On calcule $\hat{\epsilon}_i = Z(t_i) - \hat{X}(t_i)$, pour obtenir la série temporelle $\mathcal{E} = \{\epsilon_i\}_{i=1}^{T}$. Ensuite, on applique une transformation sur $\mathcal{E}$ :

$$\tilde{\mathcal{E}} = LeadLag(CS(BP(\mathcal{E}))). \tag{1}$$

Notons qu'il y a $2T + 2$ éléments dans $\tilde{\mathcal{E}}$, et chacun de ces éléments est dans $\mathbb{R}^2$.

Ensuite, on utilise une interpolation linéaire des éléments de $\tilde{\mathcal{E}}$ pour obtenir une fonction à valeurs dans $\mathbb{R}^2$. Finalement, on calcule la signature de cette fonction linéaire par morceaux. Il existe une implémentation de tout ce processus dans la fonction *SignatureTransformer* de la bibliothèque Python *sktime*, celle-ci a été dévéloppé par Morril, J., Fermanian, A., Kidger, P. et Lyons, T. J. (2020).

## 1.2 Modèle ordinal

Le modèle ordinal cumulatif classique, présenté dans le livre d'Agresti, A. (2010), permet d'étudier le lien entre une réponse ordinale $Y \in [\![1, K]\!]$ et une covariable $W \in \mathbb{R}^p$. Ce modèle s'écrit de la façon suivante :

$$logit(\mathbb{P}(Y \leq j)) = \tau_j - \langle \boldsymbol{b}, \boldsymbol{w} \rangle_{\mathbb{R}^p} \quad ; \quad [\![1, K-1]\!], \tag{2}$$

où $-\infty = \tau_0 < \tau_1 < \ldots < \tau_{K-1} < \tau_K = \infty$. Les paramètres de ce modèle sont les seuils ordonnés $(\tau_j)_{j=1}^{K-1}$ et le vecteur de paramètres $b \in \mathbb{R}^p$. Pour simplifier les notations, on désigne $\Delta_{K-1}$ l'ensemble des seuils ordonnés :

$$\Delta_{K-1} = \{(x_1, x_2, \ldots, x_{K-1}) \in \mathbb{R}^{K-1} | x_1 < x_2 < \ldots < x_{K-1}\}.$$

Pour un échantillon $\mathcal{D}$, avec $n$ observations, $\mathcal{D} = \{(W_1, Y_1), \ldots, (W_n, Y_n)\}$, l'équation (2), définit une vraisemblance $\mathcal{L}_{\mathbb{R}^p}(.|\mathcal{D})$ sur l'espace des paramètres du modèle : $\mathbb{R}^p \times \Delta_{K-1}$. Ces paramètres peuvent s'estimer par maximum de vraisemblance.

Jacques, J. et Samardzić, S. (2022) ont proposé une extension de ce modèle pour utiliser une donnée fonctionnelle en tant que variable explicative. Ils définissent le modèle suivant :

$$logit(\mathbb{P}(Y \leq j)) = \tau_j - \langle \beta, X \rangle_{L^2([0,T])} \quad ; \quad [\![1, K-1]\!]. \tag{3}$$

Notons qu'on a remplacé le produit scalaire dans $\mathbb{R}^P$ par le produit scalaire dans $L^2([0,T])$.

Pour un échantillon $\tilde{\mathcal{D}}$, avec $n$ observations, $\tilde{\mathcal{D}} = \{(X_1(.), Y_1), \ldots, (X_n(.), Y_n)\}$, l'équation (3), définit une vraisemblance $\mathcal{L}_{L^2([0,T])}(.|\tilde{\mathcal{D}})$ sur l'espace des paramètres du modèle : $L^2([0,T]) \times \Delta_{K-1}$. Afin d'estimer les paramètres, une hypothèse supplémentaire est faite sur la forme de la fonction $\beta(.)$, on suppose qu'elle peut s'écrire de la façon suivante :

$$\beta(t) = \sum_{j=1}^{P} b_j \phi(t) = \boldsymbol{b^T} \cdot \boldsymbol{\phi(t)},$$

où $(\phi_k)_{j=1}^{P}$ est une base fonctionnelle connue, $\boldsymbol{a} = (b_1, \ldots, b_P)^T$ est un vecteur de coefficients et $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_P(t))$ est un vecteur de fonctions. Sous cette hypothèse, on peut calculer un estimateur,$\tilde{\theta}_{ML}$, des paramètres du modèle par maximum de vraisemblance. Alternativement, on peut calculer un estimateur, $\tilde{\theta}_{RP}$, en maximisant la vraisemblance et contrôlant la rugosité de la fonction $\hat{\beta}$ estimée. Concrètement, on définit :

$$\hat{\theta}_{ML} = \underset{\theta \in L^2([0,T]) \times \Delta_{K-1}}{\arg\max} \mathcal{L}_{L^2([0,T])}(\theta|\tilde{\mathcal{D}}), \tag{4}$$

$$\hat{\theta}_{RP} = \underset{\theta \in L^2([0,T]) \times \Delta_{K-1}}{\arg\max} \mathcal{L}_{L^2([0,T])}(\theta|\tilde{\mathcal{D}}) + \lambda \int_{[0,T]} (\beta^{(k)}(s))^2 ds, \tag{5}$$

où $\beta^{(k)}(.)$ est la dérivée d'ordre $k$ de $\beta(.)$ et $\lambda > 0$ est un hyperparamètre connu qui contrôle la pénalisation.

Dans les deux cas, on montre dans la section suivante que ces deux problèmes d'optimisation sur un espace fonctionnel, peuvent se réécrire comme des problèmes d'optimisation dans $\mathbb{R}^P$.

## 1.3 Estimation des paramètres du modèle fonctionnel

Notons que l'on peut réécrire le produit scalaire fonctionnel comme un produit vectoriel en $\mathbb{R}^P$ :

$$\langle X, \beta \rangle_{L^2([0,T])} = \sum_{i=1}^{K} \sum_{j=1}^{P} a_i \Big( \int_{[0,T]} \psi_i(s)\phi_j(s)ds \Big) b_j = \boldsymbol{a' \cdot R \cdot b} = \langle \boldsymbol{\tilde{a}, b} \rangle_{\mathbb{R}^P},$$

où $\boldsymbol{a} = (a_1, a_2, \ldots, a_K)^T$, $\boldsymbol{b} = (b_1, b_2, \ldots, b_P)^T$, $\boldsymbol{R}_{i,j} = \int_{[0,T]} \psi_i(s)\phi_j(s)ds$ et $\boldsymbol{\tilde{a}} = \boldsymbol{a^T \cdot R}$.

En utilisant cette remarque, on peut calculer l'estimateur $\hat{\theta}_{ML}$, en estimant les paramètres d'un modèle ordinal classique. En effet, supposons que l'on observe un échantillon avec des données fonctionnelles : $\tilde{\mathcal{D}} = \{(X_i(.), Y_i)\}_{i=1}^n$ et que chaque $X_i(.)$ s'écrit $X_i(.) = \boldsymbol{a_i^T \cdot \psi}(.)$. Considérons la vraisemblance d'un modèle ordinal à covariables dans $\mathbb{R}^P$ sur l'échantillon suivant : $\mathcal{D} = \{\boldsymbol{a_i^T \cdot R}, Y_i)\}_{i=1}^n$. Alors, on a l'égalité suivante :

$$\hat{\theta}_{ML} = \underset{\theta \in L^2([0,T]) \times \Delta_{K-1}}{\arg\max} \mathcal{L}_{L^2([0,T])}(\theta | \tilde{\mathcal{D}}) = \underset{\theta \in \mathbb{R}^P \times \Delta_{K-1}}{\arg\max} \mathcal{L}_{\mathbb{R}^P}(\theta | \mathcal{D}).$$

Similairement, on peut estimer $\hat{\theta}_{RP}$ en estimant les paramètres d'un modèle ordinal, à covariables dans $\mathbb{R}^P$, avec une pénalisation ridge. On considère la matrice $P \times P$, $\boldsymbol{R}^{(k)}$, où $\boldsymbol{R}^{(k)}_{i,j} = \int_{[0,T]} \psi_i^{(k)}(s)\psi_j^{(k)}(s)ds$. Notons que :

$$\int_{[0,T]} (\beta^{(k)}(s))^2 ds = \sum_{i=1}^{P} \sum_{j=1}^{P} \boldsymbol{b}_i \left( \int_{[0,T]} \psi_i^{(k)}(s)\psi_j^{(k)}(s)ds \right) \boldsymbol{b}_j = \boldsymbol{b' R^{(k)} b}.$$

Cette matrice est symétrique positive, donc elle est diagonalisable dans une base orthonormée, on peut l'écrire de la façon suivante : $\boldsymbol{R^{(k)} = PDP'}$, où $\boldsymbol{P}$ est une matrice orthonormale et $\boldsymbol{D}$ est une matrice diagonale avec les valeurs propres, $\lambda_1, \ldots, \lambda_P$, de $\boldsymbol{R^{(k)}}$. Si on pose, $\boldsymbol{B = P^T \cdot b}$ et $\boldsymbol{\tilde{\tilde{x}} = P^T R^T a}$ alors, on a :

$$\int_{[0,T]} (\beta^{(k)}(s))^2 ds = \boldsymbol{b' R^{(k)} b} = \boldsymbol{(P^T b)^T D (P^T b)} = \boldsymbol{B^T D B} = \sum_{i=1}^{P} \lambda_i \boldsymbol{B}_i^2,$$

et

$$\langle X, \beta \rangle_{L^2([0,T])} = \boldsymbol{a^T R b} = \boldsymbol{a^T R P B} = \langle \boldsymbol{\tilde{\tilde{x}}, B} \rangle_{\mathbb{R}^P}.$$

Ainsi, si on considère l'échantillon $\tilde{\tilde{\mathcal{D}}} = \{\boldsymbol{P^T \cdot R^T a_i}, Y_i)\}_{i=1}^n$, on a :

$$\mathcal{L}_{L^2([0,T])}(\theta | \tilde{\mathcal{D}}) + \lambda \int_{[0,T]} (\beta^{(k)}(s))^2 ds = \mathcal{L}_{\mathbb{R}^P}(\theta | \tilde{\tilde{\mathcal{D}}}) + \lambda \sum_{i=1}^{P} \lambda_i \boldsymbol{B}_i^2.$$

On reconnait un modèle ordinal à covariables dans $\mathbb{R}^P$ avec une pénalité Ridge, pondérée par $\lambda_1, \ldots, \lambda_P$. Wurm, M. J., Rathouz, P. J., et Hanlon, B. M. (2021) ont développé des méthodes permettant d'estimer le modèle ordinal à covariables multivariés, avec une pénalisation ridge. Ainsi, on peut calculer $\hat{\theta}_{RP}$ avec ces méthodes.

# 2 Expériences numériques

## 2.1 Données Cookies

Afin d'illustrer la régularisation du modèle ordinal à covariables fonctionnelles, on utilise le jeu de données « Cookies ». Celui-ci a été utilisé dans les articles de Costanzo, G., Preda, C. et Saporta, G. (2006) et de Jacques, J. et Samardzić, S. (2022). Ce jeu de données provient du Danone Vitapole Paris Research Center et contient des mesures de résistance lors du pétrissage de 115 pâtes à cookies. Toutes les deux secondes, on mesure la résistance, cela pendant huit minutes, et on cuit chaque pâte pour obtenir des cookies. Ensuite, on va mesurer la qualité du cookie et voir s'il est « bon », « ajustable » ou « mauvais ». On s'intéresse au lien entre la résistance de la pâte lors du pétrissage et la qualité finale du cookie. On est clairement dans le cadre décrit dans cet article : la covariable est de nature fonctionnelle et la variable cible est de nature ordinale : mauvais <ajustable <bon.

On utilise une base de 22 splines cubiques pour représenter la résistance en tant que donnée fonctionnelle, on utilise une base de 12 splines d'ordre 5 pour représenter la fonction $\beta(.)$ et on pénalise la troisième dérivée de ce paramètre fonctionnel. Ces valeurs sont des hyperparamètres et ils sont arbitrairement choisis dans ce document. Dans la pratique, ils pourraient être sélectionnés par validation croisée.

La figure 1 permet de visualiser la résistance lors du pétrissage, selon la qualité du cookie obtenue. On s'aperçoit déjà que les bons cookies semblent avoir une résistance plus importante que les mauvais cookies, surtout vers la fin du processus. Mais ce n'est pas évident de faire cela de façon précise, une des raisons d'utiliser un modèle de régression est de quantifier cette intuition avec un modèle statistique.

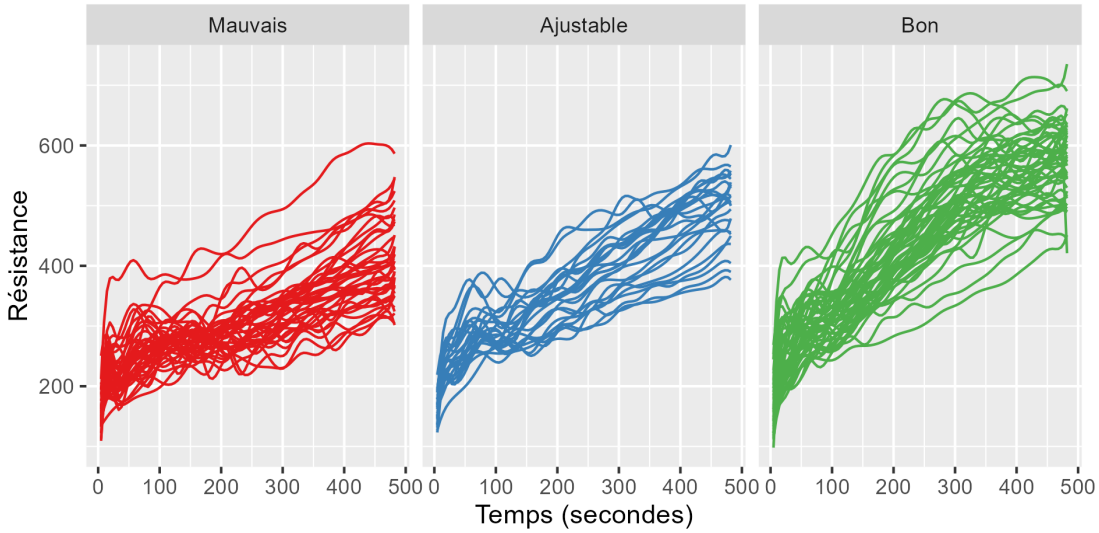



FIGURE 1 – Donnée fonctionnelle : résistance lors du pétrissage, selon qualité du cookie

Un problème pratique qui n'a pas été adressé dans la section précédente est la sélection de l'hyperparamètre $\lambda$. Une approche possible pour déterminer cet hyperparamètre est d'utiliser

une validation croisée. L'irrégularité de l'estimation de la fonction, $\beta(.)$ est un signe de sur-apprentissage, on contrôle cela en choisissant le $\lambda$ qui maximise la vraisemblance moyenne avec une validation croisée (voir la figure 2). Notons que la vraisemblance semble atteindre un maximum dans une zone (entre $\exp(-5)$ et $\exp(1)$), les estimations que l'on obtient si on pénalise trop ou pas assez donnent une vraisemblance inférieure. Notons aussi que c'est avec ce paramètre que l'on obtient le taux d'erreur de classification le plus faible.

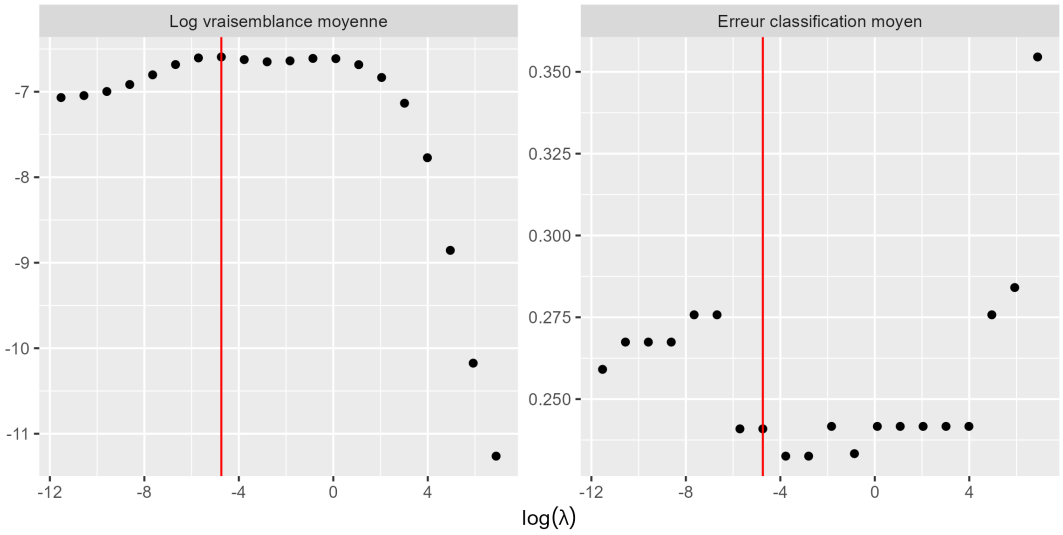



FIGURE 2 – Sélection du paramètre de l'hyperparamètre $\lambda$ par validation croisée (k=10 blocs)

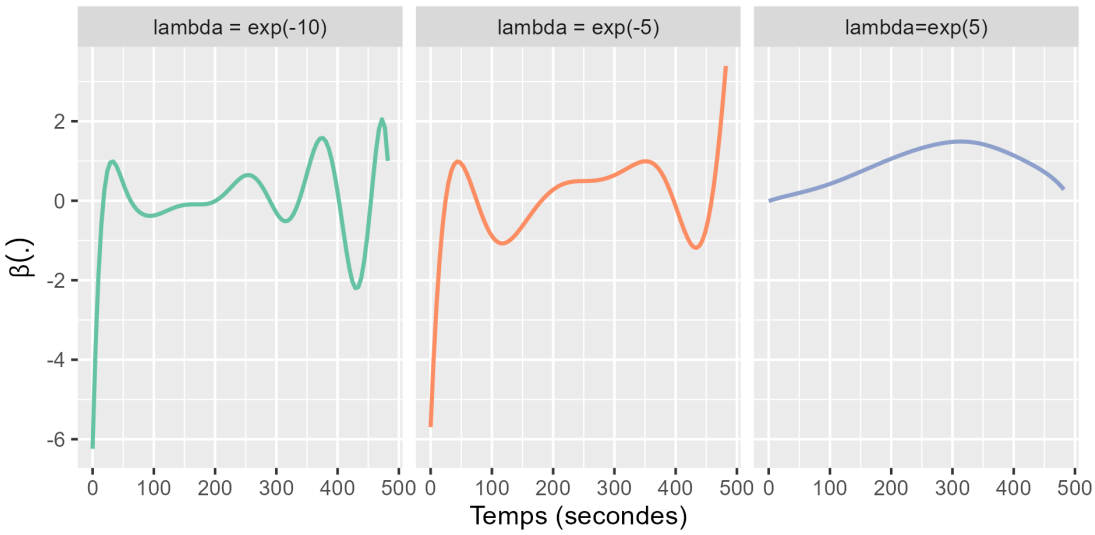



FIGURE 3 – Différentes estimations de la fonction $\beta$, selon les valeurs de $\lambda$

Finalement, notons que la forme des estimations de la fonction $\beta$ n'est pas la même selon le dégréé de régularisation (voir la figure 3). Avec un paramètre $\lambda$ trop élevé, on semble perdre de l'information temporelle, si on ne pénalise pas assez, la fonction $\beta$ estimée est difficile

à interpréter dû aux fortes oscillations (surtout entre 300 et 500 secondes). En observant l'estimation obtenue avec $\lambda = \exp(-5)$, on dirait que pour avoir un cookie de bonne qualité, il faut avoir une pâte avec une faible résistance au début et une résistance élevée au milieu et surtout à la fin du pétrissage.

## 2.2 Données BabyECG

On s'intéresse au jeu de données « BabyECG ». Ce sont des observations de l'état de sommeil d'un bébé et de son rythme cardiaque. Une mesure est enregistrée toutes les 16 secondes et cela est fait entre 21h18 et 6h27. Cette série temporelle a été présentée par Nason, G. P., Von Sachs, R. et Kroisandt, G. (2000) comme un exemple de "locally stationnary wavelet process" dont le "evolutionary wavelet spectrum" du rythme cardiaque était lié à l'état de sommeil de l'enfant. Dans la figure 4 (reproduite de l'article de Nason, G. P., Von Sachs, R. et Kroisandt, G. (2000)), on peut apercevoir ce lien.

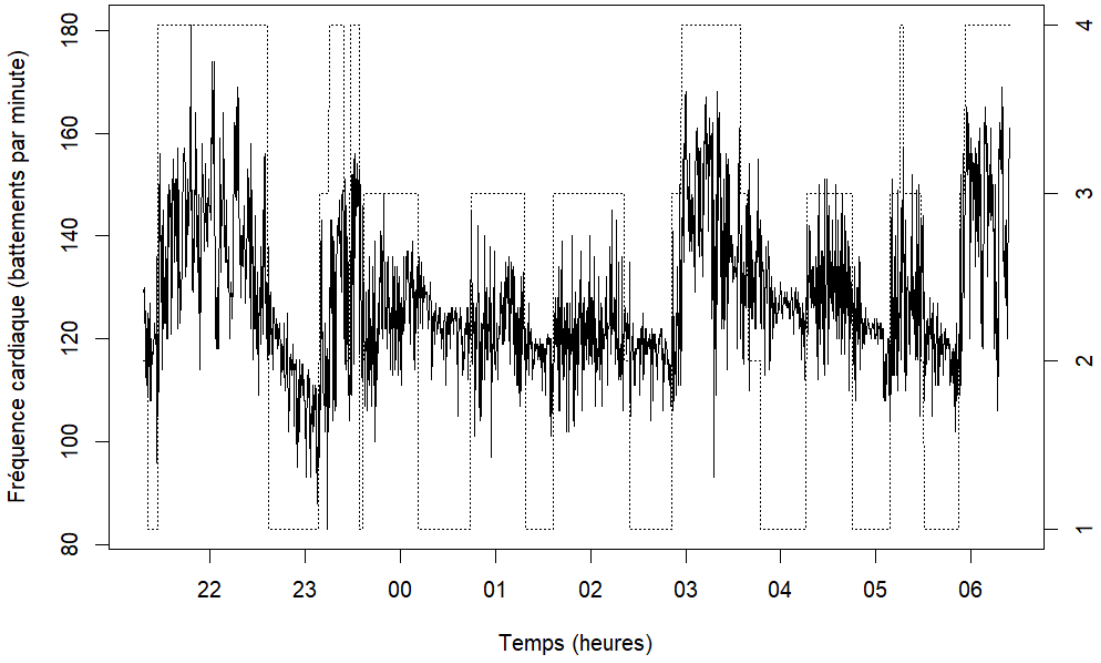



FIGURE 4 – Données BabyECG : rythme cardiaque dans le temps (ligne solide) et type de sommeil (ligne en pointillé) : les niveaux 1, 2, 3 et 4 correspondent à "réveillé", "sommeil léger", "sommeil moyen" et "sommeil profond" respectivement.

On étudie le lien entre le rythme cardiaque et le type de sommeil en utilisant la méthode de la signature présentée dans la section précédente. Pour cela, on enlève la tendance de la série entière, on divise la série obtenue dans 128 sous-séries avec 16 observations (correspondant

à 256 secondes de mesures). Pour chaque sous-série, on cherche à prédire l'état de sommeil final. On peut visualiser les sous séries obtenues dans la figure 5**A**.

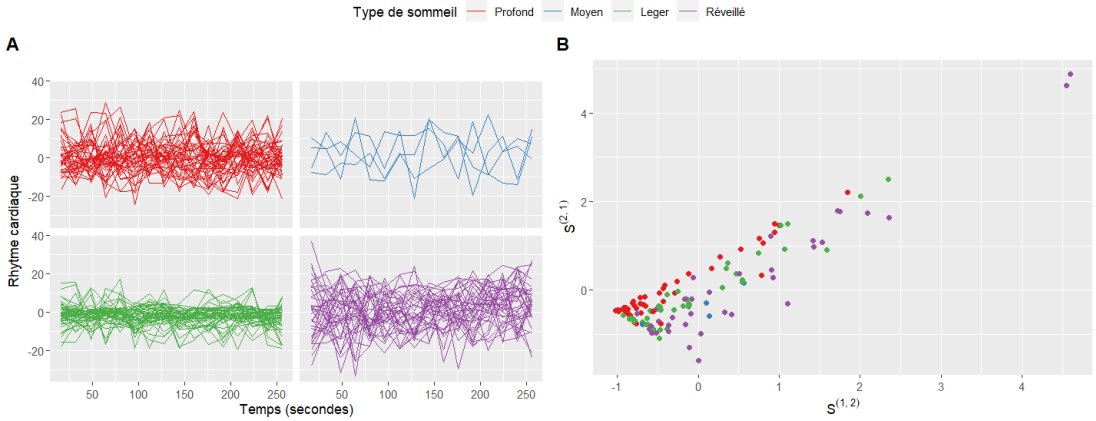



FIGURE 5 – **A** : sous-séries du rythme cardiaque obtenues, selon le type de sommeil. **B** : paire de signatures signifiantes, sélectionnées avec une pénalisation Lasso

Ensuite, pour chaque sous série de résidus, on calcule la transformation (1) et on calcule la signature de niveau deux de cette sous série bivarié. On obtient ainsi des coefficients $(S^{(1)}, S^{(2)}, S^{(1,1)}, S^{(1,2)}, S^{(2,1)}, S^{(2,2)})$, que l'on utilise comme caractéristiques pour une régression ordinale avec une pénalité lasso.

En utilisant une validation croisée avec cinq blocs, on obtient un taux d'erreur de classification de 31,3%. Les résidus de cette série apportent bien une information sur le type de sommeil et une méthode d'utiliser cette information est d'utiliser la méthode de signature. Dans la figure 5**B**, on visualise deux coefficients, $S^{(1,2)}$ et $S^{(2,1)}$, sélectionnés par Lasso. On observe que les points sont, plus ou moins, séparables par des droites parallèles. Et, sauf pour le sommeil "Moyen", les types de sommeil semblent ordonnés : on passe d'une catégorie à la catégorie adjacente.

# Conclusion

On a proposé des méthodes pour utiliser de l'information temporelle. Utiliser le modèle ordinal à covariables fonctionnelles avec la régularisation présentée, semble améliorer la qualité des prédictions obtenues et ça facilite l'interprétation du modèle. Utiliser la méthode basée dans des signatures permet d'extraire de l'information qui serait perdue avec un modèle à covariables fonctionnelles.

Il est de plus en plus usuel d'avoir des objets équipés de capteurs. De tels objets permettent d'obtenir des mesures successives qui peuvent être traitées comme des variables qui dépendent du temps. Les approches proposées dans cet article permettraient de se servir de telles variables.

# Bibliographie


Agresti, A. (2010), *Analysis of Ordinal Categorical Data*, Wiley Series in Probability and Statistics.

Chevyrev, I. et Kormilitzin, A. (2016), A primer on the signature method in machine learning, *arXiv preprint arXiv :1603.03788*.

Costanzo, G., Preda, C. et Saporta, G. (2006), Anticipated prediction in discriminant analysis on functional data for binary response, *COMPSTAT'06, 17th Symposium on Computational Statistics*.

Fermanian, A. (2021), *Learning time-dependent data with the signature transform*, Thèse Université Sorbonne.

Jacques, J. et Samardzić, S. (2022), Analyzing cycling sensors data through ordinal logistic regression with functional covariates, *Journal of the Royal Statistical Society : Series C Applied Statistics*.

Morril, J., Fermanian, A., Kidger, P. et Lyons, T. J. (2020), A generalised signature method for time series, *arXiv preprint arXiv :2006.00873*.

Nason, G. P., Von Sachs, R. et Kroisandt, G. (2000), Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62, pp. 271–292.

Ramsay, J. O. et Silverman, B. W. (2005), *Functional data analysis*, Springer series in statistics.

Wurm, M. J., Rathouz, P. J., et Hanlon, B. M. (2021), Regularized ordinal regression and the ordinalnet r package, *Journal of Statistical Software*, 99, pp. 1-42.

# IA Générative

# Conditional denoising diffusion probabilistic models for the clustering of images

Seydina NIANG [1] & Charles BOUVEYRON [1] & Marco CORNELI[1,2] & Pierre LATOUCHE [3]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France*
[2] *Université Côte d'Azur, Laboratoire CEPAM, Nice, France*
[3] *Université Clermont Auvergne, CNRS, Laboratoire LMBP, Aubière, France*

**Résumé.** Dans le domaine du traitement d'images, les modèles de diffusion du débruitage (DDPM) ont gagné en popularité pour leur capacité à modéliser les distributions de données complexes tout en permettant une génération réaliste d'images. Ces modèles reposent sur deux étapes principales : une phase de diffusion directe où les images sont graduellement corrompues par du bruit gaussien, suivie d'une étape de décodage inverse où les images bruitées sont débruitées étape par étape à l'aide d'un réseau de neurones. Dans ce travail, nous proposons une extension du modèle DDPM en introduisant une version conditionnelle qui prend en compte l'appartenance des images à différents groupes. Cette approche permet de mieux capturer la structure sous-jacente des données en tenant compte de leur regroupement naturel. Concrètement, le modèle repose sur l'hypothèse que les images sont réparties en $Q$ groupes, et chaque image est modélisée comme provenant d'une distribution conditionnelle sur les groupes. Un réseau de neurones est ensuite entraîné pour prédire le bruit à supprimer lors de la génération des images, en fonction de leur appartenance à un groupe.

L'inférence dans ce modèle complexe est effectuée à l'aide d'un algorithme de type EM variationnel, qui permet d'estimer de manière efficace les paramètres de regroupement et les variables latentes. Cependant, l'optimisation des paramètres du réseau de neurones pose des défis supplémentaires, nécessitant l'utilisation de techniques avancées comme la descente de gradient stochastique. En combinant les avantages des modèles de diffusion du débruitage avec une approche de clustering conditionnelle, cette méthode ouvre de nouvelles perspectives dans le domaine de l'apprentissage automatique et du traitement d'images, offrant des outils puissants pour l'analyse et la génération de données visuelles complexes.
**Mots-clés.** Modèles de diffusion du débruitage, traitement d'images, algorithme EM, descente de gradient stochastique, inférence variationnelle.

**Abstract.** This document presents an innovative method for clustering image data while learning to generate new images. In the field of image processing, denoising diffusion probabilistic models (DDPMs) have gained popularity for their ability to model complex data distributions while enabling realistic image generation. These models rely on two main steps: a forward diffusion phase where images are gradually corrupted by Gaussian noise, followed by a reverse decoding step where the noisy images are denoised step by step using a neural network. In this work, we propose an extension of the DDPM model by introducing a conditional version that takes into account the membership of images to different clusters. This approach allows for better capturing the underlying structure of the data by considering their natural grouping. Specifically, the model is based on the

assumption that images are distributed into $Q$ clusters and that each image is modeled as coming from a conditional distribution over the clusters. A neural network is then trained to predict the noise to be removed during data generation, based on the image membership to a specific cluster. Inference in this complex model is performed using a variational EM-type algorithm, which efficiently estimates the clustering parameters and latent variables. However, optimizing the parameters of the neural network poses additional challenges, requiring the use of techniques such as stochastic gradient descent.
By combining the advantages of denoising diffusion models with a conditional clustering approach, this method opens up new perspectives in the field of machine learning and image processing, providing powerful tools for the analysis and generation of complex visual data.
**Keywords.** Diffusion probabilistic models, variational inference, expectation-maximisation algorithm (EM), image processing, stochastic gradient descent.

# 1 Introduction

In recent years, generative models have made significant strides in generating human-like natural language, high-quality synthetic images and diverse human speech and music. These models find applications in various domains, such as generating images from text prompts or learning useful feature representations. Despite their ability to produce realistic outputs, there remains ample room for improvement in generative models, which could have broad implications across graphic design, gaming, music production and beyond.
Generative adversarial networks (GAN) (Creswell et al., 2018) recently led the field of image generation tasks, as measured by metrics like FID (Fréchet inception distance), inception score and precision used to evaluate the quality and diversity of generated images. However, GANs often struggle with diversity and can be challenging to train effectively, requiring careful tuning of hyper-parameters and regularizers. While likelihood-based models offer advantages in terms of diversity and ease of training, they still fall short in visual fidelity compared to GANs. Diffusion models, a class of likelihood-based models, have shown promising results in producing high-quality images with desirable properties such as distribution coverage and scalability. As shown in Dhariwal and Nichol, 2021, they can outperform GANs in the context of image processing. Here we propose to condition the diffusion process to the assignment to a cluster of the images and we believe that this can improve the efficiency of the generation while learning a way to cluster the images.

# 2 Conditional diffusion probabilistic model for the clustering of images

## 2.1 Denoising diffusion probabilistic models

A denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel, 2020) makes use of two Markov chains: a forward chain that perturbs data to noise and a reverse chain that converts noise back to data. The former is typically hand-designed with the goal

to transform any data distribution into a simple prior distribution (e.g., standard Gaussian), while the latter Markov chain reverses the former by learning transition kernels parameterized by deep neural networks. New data points are subsequently generated by first sampling a random vector from the prior distribution, followed by ancestral sampling through the reverse Markov chain. Formally, given a target data distribution $x_0 \sim q^\star(x_0)$, the forward Markov process generates a sequence of random variables $x_1, x_2 \ldots, x_T$ with transition kernel $q(x_t|x_{t-1})$. Using the chain rule of probability and the Markov property, we can factorize the joint distribution of $x_1, x_2, \ldots x_T$ conditioned on $x_0$, denoted as $q(x_1, \ldots, x_T|x_0)$, into

$$q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{1}$$

In DDPMs, we handcraft the transition kernel $q(x_t|x_{t-1})$ to incrementally transform the data distribution $q^\star(x_0)$ into a tractable prior distribution. One typical design for the transition kernel is Gaussian perturbation and the most common choice for the transition kernel is

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \cdot \mathbf{I}_D), \tag{2}$$

where $D$ is the dimensionality of the data $x_0$ and $\beta_t \in (0, 1)$ is a hyper-parameter chosen ahead of model training. This kernel is the most used, although other types of kernels are also applicable in the same vein. Specifically, with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we have

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t) \cdot \mathbf{I}_D). \tag{3}$$

Given $x_0$, we can easily obtain a sample of $x_t$ by sampling a Gaussian vector $\epsilon \sim \mathcal{N}(0, \mathbf{I}_D)$ and applying the reprametrisation trick:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon. \tag{4}$$

When $\bar{\alpha}_T \sim 0$ (true for high values for $T$), $x_T$ follows a Gaussian white noise $q(x_T) \simeq \mathcal{N}(X_T; 0, \mathbf{I}_D)$. Intuitively, this forward process slowly injects noise to data until all structures are lost.

For generating new data samples, DDPMs start by first generating an unstructured noise vector from the prior distribution (which is typically trivial to obtain), then gradually remove noise by running a learnable Markov chain in the reverse time direction. Specifically, the reverse generative Markov chain is parameterized by a prior distribution $p(x_T) = \mathcal{N}(X_T; 0, \mathbf{I}_D)$ and a learnable transition kernel $p_\theta(x_{t-1}|x_t)$. This learnable transition kernel takes the form of

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{5}$$

where $\theta$ denotes model parameters and the mean $\mu_\theta(x_t, t)$ and variance $\Sigma_\theta(x_t, t)$ are parameterized by a deep neural networks. With this reverse Markov chain in hand, we can generate a data sample $x_0$ by first sampling a noise vector $x_T \sim p(x_T)$, then iteratively sampling from the learnable transition kernel $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ until $t = 1$.

Key to the success of this sampling process is training the reverse Markov chain to match the actual time reversal of the forward Markov chain. That is, we have to adjust the parameter $\theta$ so that the joint distribution of the reverse Markov chain $p_\theta(x_0, \ldots x_T)$ closely approximates that of the forward process $q(x_0, \ldots x_T)$. This is achieved by minimizing the Kullback-Leibler (KL) divergence between these two:

$$\mathcal{L} = \mathrm{KL}(p_\theta(x_0, \ldots x_T)||q(x_0, \ldots x_T)). \tag{6}$$

To sum up, diffusion models are powerful generative models that work by gradually adding noise to a latent representation of the data and then inverting the process to reconstruct the data. This process can be used to create realistic images, even when the latent representation is simple. A major question is whether diffusion models are flexible enough to be adapted to the clustering field or not.

## 2.2 Our contribution: mixture of denoising diffusion probabilistic models

Mixture models are a powerful statistical technique used in machine learning and data analysis, particularly for the purpose of clustering. Clustering is the process of grouping similar data points together, where the similarity between data points is defined based on certain characteristics or features. Through this section, we build a mixture of denoising diffusion probabilistic models for the aim of images clustering.
Let us suppose that the train dataset is gathered into a matrix $\{X_i^0\}_{i\leq N}$ corresponding to a collection of images regrouped into $Q$ types (clusters). Each line $X_i^0$ of the matrix $X_0$ is a flattened image in $R^D$. $N$ denotes the number of images considered and $D$ is the number of pixels.

**Generative model:**
We posit the following distribution for the data:

$$p(X^0|\mu,\theta,\pi) = \prod_{i=1}^{N}\sum_{q=1}^{Q}\pi_q p_\theta(X_i^0|\mu_q), \tag{7}$$

where $\pi = (\pi_q)_q$ denotes the mixture proportions (the probability that an image is drawn following the $q$-th component of the mixture), $\mu_q$ corresponds to the mean of the prior Gaussian distribution of the $q$-th diffusion model and $p_\theta$ refers to a conditional diffusion model (Lu et al., 2022). Indeed we suppose that the diffusion process (the way noise is added to images) is specific to each cluster. Clustering arises from this probabilistic formulation with the introduction of an unobserved random variable $Z_i \in \{0,1\}^Q$ such that $Z_{iq} = 1$ if and only if the image $i$ belongs to the $q$-th component. $Z_i$ is supposed to be drawn following a multinomial distribution of parameter $\pi$

$$Z_i \overset{\text{iid}}{\sim} \text{Multinomial}(1, \pi = (\pi_{1:Q})), \quad \forall i \in \{1,\dots,N\}. \tag{8}$$

The corresponding partition $Z = \{Z_i\}$ is then considered as the set of discrete latent variables which are to be estimated, along with the model parameters. Traditionally the posterior distribution of $Z$ is estimated by an expectation-maximization algorithm (EM) through variational inference.
Hence the conditional distribution of the $i$-th image given $Z_i$ is:

$$X_{i\,|Z_{iq}=1}^0 \sim p_\theta(\cdot|\mu_q).$$

Instead of generating image directly from latent noise as done in variational auto-encoders (Kipf and Welling, 2016), we construct the image sequentially. We begin by introducing for each image $i$, $T \in \mathbf{N}^\star$ latent variables $X_i^{1:T}$ such that $X_i^{0:T}$ conditioned to $Z$ is a reverse

time Markov chain with transition kernel $p_\theta(X_i^{t-1}|X_i^t, \mu_q)$. Hence:

$$p_\theta(X_i^{0:T}|Z,\mu) = \prod_{q=1}^{Q}\left(p(X_i^T)\prod_{t=1}^{T}p_\theta(X_i^{t-1}|X_i^t,\mu_q)\right)^{Z_{iq}}, \tag{9}$$

where

$$\begin{aligned} p(X_i^T) &= \mathcal{N}(X_i^T; 0, I_D) \quad \text{and} \\ p_\theta(X_i^{t-1}|X_i^t, Z_{iq}=1, \mu_q) &= \mathcal{N}(X_i^{t-1}; \mu_\theta(X_i^t, t, \mu_q), \bar{\delta}_t \cdot I_D), \end{aligned} \tag{10}$$

with $\mu_\theta$ a neural network and $\bar{\delta} > 0$ an hyper-parameter.
We point out that the reverse transition kernel now also depends on the cluster means $\mu_q$. This assumption seems natural, since we start with a white noise and want to generate images of a certain type.

# 3 Inference

The complete-data log-likelihood is computed by integrating on all the latent variables, giving:

$$\log p_\theta(X^0) = \log\left[\sum_Z \int_{X^1,\dots,X^T,} p_\theta(X^0,\dots,X^T,Z)dX^1\dots dX^T\right], \tag{11}$$

where , $X^t := (X_i^t)_i$ for all $t$ and all $i$.
Unfortunately, the above log-likelihood is untractable. To tackle this problem, we rely on variational inference. We introduce a variational posterior distribution $q(\cdot)$ on the latent variables that factorizes as:

$$\begin{aligned} q(X^1,\dots,X^T,Z|X^0) &= q(X^0,\dots,X^T|Z,X^0)\cdot q(Z) \\ &= \prod_{i=1}^{N} q_i(Z_i)\prod_{q=1}^{Q}\left(\prod_{t=1}^{T} q(X_i^t|X_i^{t-1},\mu_q)\right)^{Z_{iq}}, \end{aligned} \tag{12}$$

where

$$q_i(Z_i) := \mathcal{M}(Z_i; 1, \tau_i), \tag{13}$$

and

$$q(X_i^t|X_i^{t-1},\mu_q) = \mathcal{N}\left(X_i^t; \frac{1-m_t}{1-m_{t-1}}\sqrt{\alpha_t}X_i^{t-1} + \left(m_t - \frac{1-m_t}{1-m_{t-1}}m_{t-1}\right)\sqrt{\bar{\alpha}_t}\mu_q, \quad \delta_{t|t-1}I_D\right), \tag{14}$$

with $\delta_{t|t-1} = \delta_t - \left(\frac{1-m_t}{1-m_{t-1}}\right)^2$, $\delta_t = (1-\bar{\alpha}_t)^2 - m_t^2\bar{\alpha}_t$ and $m_0 \simeq 0 \leq \dots \leq m_T \simeq 1$.
As it can be seen, the model is flexible enough to introduce noise in distinct ways based on cluster membership. What sets our approach apart is the unconventional aspect of both the variational and generative distributions being governed by the same parameter $\mu$. However we need both the forward and reverse process (i.e. from image to noise) to be cluster-dependant.

$q(X_i^t|X_i^0,\mu_q)$ is computed by marginalisation w.r.t the intermediate latent variables, resulting into

$$q(X_i^t|X_i^0, Z_{iq}=1, \mu_q) = \mathcal{N}(X_i^t; (1-m_t)\sqrt{\bar{\alpha}_t}X_i^0 + m_t\sqrt{\bar{\alpha}_t}\mu_q, \delta_t I_D). \tag{15}$$

Here the mean of the Gaussian distribution is an interpolation of the original image and the cluster mean. The interest of this formula can be outlined by the reparametrization trick:

$$X_i^T = (1 - m_T)\sqrt{\bar{\alpha}_T}X_i^0 + m_T\sqrt{\bar{\alpha}_T}\mu_q + \sqrt{\delta_t}\eta_T,$$

where $\eta_T \sim \mathcal{N}(0, I_D)$.
Given that $m_T \simeq 1$ and $\alpha_T \to 0$, it follows that $X_i^T$ behaves nearly like white noise, denoted as $X_i^T \simeq \eta_T$. This result is coherent with $p(X_i^T) = \mathcal{N}(X_i^T; 0, I_D)$.

If we inject the variational distribution into the untractatable log-likelihood, we obtain:

$$\begin{aligned}\log p_\theta(X^0) &= \log\left[\sum_Z \int_{X^1,\dots,X^T} \frac{p_\theta(X^0,\dots,X^T,Z)}{q(X^1,\dots,X^T,Z|X^0)} \cdot q(X^1,\dots,X^T,Z|X^0)dX^1\dots dX^T\right] \\ &= \log E_{X^1,\dots,X^T,Z\sim q(\cdot|X^0)}\left[\frac{p_\theta(X^0,\dots,X^T,Z)}{q(X^1,\dots,X^T,Z|X^0)}\right].\end{aligned} \tag{16}$$

Since log is concave, the Jensen inequality gives:

$$\log p_\theta(X^0) \geq E_{X^1,\dots,X^T,Z\sim q(\cdot)}\left[\log\frac{p_\theta(X^0,\dots,X^T,Z)}{q(X^1,\dots X^T,Z|X^0)}\right] =: \mathcal{L}(\theta,\mu,\tau,\pi). \tag{17}$$

We have derived one variational lower bound $\mathcal{L}(\cdot)$ from the marginal log-likelihood, we now proceed by optimising the lower bound w.r.t. the parameters.

**Optimisation**

In this subsection, we delve into the optimization process for our model. We start by presenting a theorem that decomposes the variational lower bound.

**Proposition 1.** *The variational lower bound can be decomposed as following:*

$$\begin{aligned}\mathcal{L}(\cdot) = &\sum_{i,q}\tau_{iq}\left[-\frac{1}{2}\left|\left|\sqrt{\bar{\alpha}_T}\mu_q\right|\right|_2^2 - \sum_{t>1}E_{X^t}\left[\frac{1}{2\bar{\delta}_t}(\left|\left|\tilde{\mu}(X_i^t,X_i^0,\mu_q) - \mu_\theta(X_i^t,t,\mu_q)\right|\right|_2^2)\right] + \log\pi_q - \log\tau_{i,q}\right] \\ &+ \sum_{i,q}\tau_{iq}E_{X^1}\log p_\theta(X_i^0|X_i^1,\mu_q)\end{aligned} \tag{18}$$

*where*

$$\begin{aligned}&E_{X^t} \triangleq E_{X^t\sim q} \quad \textit{and} \\ \tilde{\mu}(X_i^t,X_i^0,\mu_q) = &\frac{1-m_t}{1-m_{t-1}}\frac{\delta_{t-1}}{\delta_t}\sqrt{\alpha_t}X_i^t + (1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}\sqrt{\bar{\alpha}_{t-1}}X_i^0 \\ &+ \left(m_{t-1}\delta_t - \frac{m_t(1-m_t)}{1-m_{t-1}}\alpha_t\delta_{t-1}\right)\frac{\sqrt{\bar{\alpha}_{t-1}}}{\delta_t}\mu_q.\end{aligned} \tag{19}$$

Note that by setting $\beta_1$ to a very small value,one has $X_i^1 \sim X_i^0$ so that the last term of $\mathcal{L}$ can be neglected.
Moreover, since $X_i^t|X_i^0 \sim \mathcal{N}((1-m_t)\sqrt{\bar{\alpha}_t}X_i^0 + m_t\sqrt{\bar{\alpha}_t}\mu_q; \delta_t I_D)$, then we can re-parameterise $X_i^t$ as:

$$X_i^t = (1-m_t)\sqrt{\bar{\alpha}_t}X_i^0 + m_t\sqrt{\bar{\alpha}_t}\mu_q + \sqrt{\delta_t}\epsilon_t \quad \text{where } \epsilon_t \sim \mathcal{N}(0, I_D).$$

Then any expectation with respect to $X_i^t$ can be transformed to an expectation with respect to the couple $(X_i^0, \epsilon_t)$ (the proof of this can easily be obtained by a change of variable on the integral).
Now we have to fix an architecture for the neural network $\mu_\theta$.

**Architecture of $\mu_\theta$**:
Notice that the optimal value for $\mu_\theta(X_i^t, t, \mu_q)$ is $\tilde{\mu}(X_i^t, X_i^t, \mu_q)$ (in Eq 19)

As $\mu_\theta$ allows us to remove the correct amount of noise to pass from latent image $X_i^t$ to $X_i^{t-1}$ and given that this later was added linearly in the forward process (see the functional form of $q(X_i^t|X_i^{t-1})$), we suppose that $\mu_\theta$ has the following functional form:

$$\mu_\theta(X_i^t, t, \mu_q) = c_{X^t} X_i^t + c_{\mu_q}\mu_q + c_{\eta_\theta}\eta_\theta(X_i^t, t, \mu_q)$$

where $\eta_\theta$ is a neural network and

$$\begin{aligned} c_{X_t} &= \frac{1-m_t}{1-m_{t-1}}\frac{\delta_{t-1}}{\delta_t}\sqrt{\alpha_t} + (1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}\frac{1}{\sqrt{\alpha_t}}. \\ c_{\mu_q} &= \left(m_{t-1}\delta_t - \frac{m_t(1-m_t)}{1-m_{t-1}}\alpha_t\delta_{t-1}\right)\frac{\sqrt{\bar{\alpha}_{t-1}}}{\delta_t}. \\ c_{\eta_t\theta} &= -(1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t\sqrt{\alpha_t}}\sqrt{1-\bar{\alpha}_t}. \end{aligned}$$

After some calculations, one gets

$$\tilde{\mu} - \mu_\theta = c_{\eta_t}\left(\frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(m_t\sqrt{\bar{\alpha}_t}(\mu_q - X_i^0) + \sqrt{\delta_t}\epsilon_t\right) - \epsilon_\theta\right). \tag{20}$$

Hence
$\mathcal{L} \simeq \sum_{i,q}\tau_{iq}\left[-\frac{1}{2}\left|\left|\sqrt{\bar{\alpha}_T}\mu_q\right|\right|_2^2 - \sum_{t>1}E_{X_i^0,\epsilon}\left[\frac{c_{\eta_t}^2}{2\delta_t}\left|\left|\frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(m_t\sqrt{\bar{\alpha}_t}(\mu_q - X_i^0) + \sqrt{\delta_t}\epsilon_t\right) - \eta_\theta\right|\right|_2^2\right] + \log\pi_q - \log\tau_{i,q}\right]$

## 3.1 Variational EM algorithm

**Proposition 2.** *The optimal updates of $\tau_{iq}$ (E-step) and $\pi_q$ (M-step) are given by:*

$$\begin{aligned} \tau_{iq} &= \frac{\pi_q \exp\left(-\frac{1}{2}||\sqrt{\bar{\alpha}_T}\mu_q||_2^2 - \sum_{t>1}\frac{1}{2\delta_t}f_{\theta,i,t,q} - 1\right)}{\sum_{q=1}^Q \pi_q \exp\left(-\frac{1}{2}||\sqrt{\bar{\alpha}_T}\mu_q||_2^2 - \sum_{t>1}\frac{1}{2\delta_t}f_{\theta,i,t,q} - 1\right)}, \\ \pi_q &= \frac{N_q}{N}, \end{aligned} \tag{22}$$

*where $N_q = \sum_i \tau_{iq}$ and*

$$\begin{aligned} f_{\theta,i,t,q} &= E_{X_i^t}\left|\left|\tilde{\mu}(X_i^t, t, \mu_q) - \mu_\theta(X_i^t, t, \mu_q)\right|\right|_2^2 \\ &= E_{X_i^0,\epsilon_t}\left[\left[\frac{c_{\eta_t}^2}{2\bar{\delta}_t}\left|\left|\frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(m_t\sqrt{\bar{\alpha}_t}(\mu_q - X^0) + \sqrt{\delta_t}\epsilon_t\right) - \epsilon_\theta(M(X_i^0, t, \epsilon_t), t, \mu_q)\right|\right|_2^2\right]\right], \end{aligned}$$

*with $M(X_i^0, t, \epsilon_t) = (1-m_t)\sqrt{\bar{\alpha}_t}X_i^0 + m_t\sqrt{\bar{\alpha}_t}\mu_q + \sqrt{\delta_t}\epsilon_t$.*

To optimize the model's parameters (M-step), we employ the VB-EM algorithm, as outlined in (Kipf and Welling, 2016), which is described in Algorithm 1.

We can see that the exact value of $\tau_{iq}$ cannot be explicitly determined due to the expectation inside $f_{\theta,i,t,q}$.

# 4 Conclusion

We have demonstrated the potential extension of diffusion models, a class of likelihood-based models with a stationary training objective, to the clustering field, akin to numerous other generative models. Unfortunately, due to time constraints and the unresolved issue regarding $\tau_{iq}$, experimental validation was not feasible within this study. However, we intend to present these experiments directly during the presentation should our paper be accepted, providing a more comprehensive evaluation of our approach.

**Algorithm 1:** Training algorithm

**while** *not convergence, for each step* $k$ **do**
  **For** $\boldsymbol{i = 1, \dots N}$
    **For** $\boldsymbol{q = 1, \dots Q}$
      Compute $\tau_{iq}^k$ and $\pi_q^k$ according to Equation 22
  **For** $\boldsymbol{l = 1, 2, \dots N_{\text{iter}}}$ **do**
    **For** $\boldsymbol{q = 1, \dots Q}$ **do**
      Sample $t \sim \text{Uniform}\{1, \dots T\}$, $\epsilon_t \sim \mathcal{N}(0, I_D)$ and $i \sim \text{Uniform}\{1, \dots N\}$
      Compute $X_i^t = (1 - m_t)\sqrt{\bar{\alpha}_t} X_i^0 + m_t \sqrt{\bar{\alpha}_t}\mu_q + \sqrt{\delta_t}\epsilon_t$
      Take gradient step on

$$\nabla_{\theta,\mu_q}\tau_{iq}\left[\left|\left|\frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(m_t\sqrt{\bar{\alpha}_t}(\mu_q - X_i^0) + \sqrt{\delta_t}\epsilon_t\right) - \eta_\theta(X_{iq}^t, t, \mu_q)\right|\right|_2^2 + \frac{\bar{\alpha}_T}{2}\,||\mu_q||_2^2\right]$$

# Bibliography

Creswell, Antonia et al. (2018). "Generative adversarial networks: An overview". In: *IEEE signal processing magazine* 35.1, pp. 53–65.

Dhariwal, Prafulla and Alexander Nichol (2021). "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34, pp. 8780–8794.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33, pp. 6840–6851.

Kipf, Thomas N and Max Welling (2016). "Variational graph auto-encoders". In: *arXiv preprint arXiv:1611.07308*.

Lu, Yen-Ju et al. (2022). "Conditional diffusion probabilistic model for speech enhancement". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7402–7406.

# Wasserstein GAN are minimax optimal distribution estimators

Arthur Stéphanovitch [1] & Eddie Aamari [2] & Clément Levrard [3]

[1] *Université Paris Cité, Sorbonne Université, CNRS*
*Laboratoire de Probabilités, Statistique et Modélisation*
*Paris, France*
[2] *Ecole Normale Supérieure, Université PSL, CNRS*
*Département de Mathématiques et Applications*
*F-75005 Paris, France*
[3] *Université de Rennes, CNRS*
*Institut de recherche mathématique de Rennes*
*F-35000 Rennes, France*

**Résumé.** Nous étudions les taux de convergence non asymptotiques de l'estimateur Wasserstein Generative Adversarial Networks (WGAN). Précisément, on construit des classes de réseaux neuronaux représentant les générateurs et les discriminateurs qui donnent un GAN atteignant le taux minimax optimal pour l'estimation d'une certaine mesure de probabilité $\mu$ avec un support dans $\mathbb{R}^p$. La probabilité $\mu$ est considérée comme la poussée en avant de la mesure de Lebesgue sur le tore $d$-dimensionnel $\mathbb{T}^d$ par une application $g^\star : \mathbb{T}^d \to \mathbb{R}^p$ de régularité $\beta + 1$. En mesurant l'erreur avec l'IPM $\gamma$-Hölder , nous obtenons, à des facteurs logarithmiques près, le taux minimax optimal $O(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}})$, où $n$ est la taille de l'échantillon, $\beta$ détermine la régularité de la mesure cible $\mu$, $\gamma$ est la régularité de l'IPM ($\gamma = 1$ dans le cas de Wasserstein), et $d \leq p$ est la dimension intrinsèque de $\mu$. Dans le processus, nous montrons une inégalité d'interpolation entre les IPM Hölder. Ce résultat de la théorie des espaces de fonctions généralise les inégalités d'interpolation classiques au cas où les mesures impliquées ont des densités sur des variétés différentes.

**Mots-clés.** Taux minimax, modèle génératif, estimation de mesure, sous-variété, inégalité d'interpolation

**Abstract.** We provide non asymptotic rates of convergence of the Wasserstein Generative Adversarial networks (WGAN) estimator. We build neural networks classes representing the generators and discriminators which yield a GAN that achieves the minimax optimal rate for estimating a certain probability measure $\mu$ with support in $\mathbb{R}^p$. The probability $\mu$ is considered to be the push forward of the Lebesgue measure on the $d$-dimensional torus $\mathbb{T}^d$ by a map $g^\star : \mathbb{T}^d \to \mathbb{R}^p$ of smoothness $\beta + 1$. Measuring the error with the $\gamma$-Hölder Integral Probability Metric (IPM), we obtain up to logarithmic factors, the minimax optimal rate $O(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}})$ where $n$ is the sample size, $\beta$ determines the smoothness of the target measure $\mu$, $\gamma$ is the smoothness of the IPM ($\gamma = 1$ is the Wasserstein case) and $d \leq p$ is the intrinsic dimension of $\mu$. In the process, we derive a sharp interpolation inequality between Hölder IPMs. This novel result of theory of functions spaces generalizes classical interpolation inequalities to the case where the measures involved have densities on different manifolds.

**Keywords.** Minimax rate, generative model, distribution estimation, manifold, interpolation inequality

# 1 Résumé long

Soient $X_1, ..., X_n$ des points aléatoires i.i.d. tirés d'une mesure de probabilité $\mu$ avec un support dans $\mathbb{R}^p$. L'inférence de $\mu$ est un problème fondamental en statistique et en apprentissage automatique, pour lequel de nombreuses méthodes ont été développées (Tsybakov, 2004). Ces dernières années ont vu l'avènement de méthodologies génératives basées sur les réseaux antagonistes génératifs (GAN) (Goodfellow et al., 2014), avec des réalisations exceptionnelles dans les domaines de l'image (Karras et al., 2021), de la vidéo (Vondrick et al., 2016), et de la génération de texte (Yu et al., 2017). Dans cet article, nous nous concentrons sur l'approche GAN de Wasserstein (WGAN) de (Arjovsky et al., 2017), qui utilise la distance de Wasserstein 1 comme alternative à la divergence de Jensen-Shannon mise en œuvre dans le GAN traditionnel. Au fil des ans, les WGAN et leurs dérivés ont gagné en popularité dans la communauté de l'apprentissage automatique. Ils sont aujourd'hui considérés comme l'une des techniques génératives les plus réussies, obtenant des résultats de pointe dans des problèmes difficiles (Karras et al., 2021), tout en améliorant la stabilité et en éliminant des problèmes désagréables tels que le collapsus de mode (Gulrajani et al., 2017). Bien que les WGAN aient montré d'excellentes propriétés dans de nombreuses études empiriques rapportées dans la littérature sur l'apprentissage automatique (Liu et al. (2019), Luo and Lu (2018), Stanczuk et al. (2021)), bon nombre de leurs propriétés théoriques restent à étudier.

Le présent travail établit l'optimalité minimax de l'estimateur WGAN. Pour mettre en place la notation, le problème génératif consiste à utiliser les données $X_1, ..., X_n$ pour apprendre $\mu$ et, simultanément, être capable d'échantillonner à partir d'une distribution proche de celle-ci. Afin de résoudre ce problème, un réseau antagoniste génératif se compose d'une classe de fonctions génératrices $\mathcal{G}$ et d'une classe de discriminateurs $\mathcal{D}$. Étant donné une distribution facile à échantillonner $\nu$ sur un espace latent $\mathcal{Z}$, le générateur $\mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R}^p$ approxime $\mu$ en essayant de minimiser sur $\mathcal{G}$ une certaine métrique de probabilité intégrale (IPM) (Müller, 1997) :

$$d_{\mathcal{D}}(\mu, g_{\#\nu}) := \sup_{D \in \mathcal{D}} \mathbb{E}_{\mu}[D(X)] - \mathbb{E}_{\nu}[D(g(Z))], \tag{1}$$

où $g_{\#\nu}$ représente la mesure de transfert de $\nu$ par $g$. L'objectif du discriminateur $\mathcal{D} \ni D : \mathbb{R}^p \to \mathbb{R}$ est de distinguer entre la vraie distribution et la fausse $g_{\#\nu}$, en maximisant sur $\mathcal{D}$ la quantité

$$L(g, D) := \mathbb{E}_{\mu}[D(X)] - \mathbb{E}_{\nu}[D(g(Z))].$$

Le problème min-max des réseaux antagonistes génératifs peut alors être écrit comme

$$\inf_{g \in \mathcal{G}} \sup_{D \in \mathcal{D}} \mathbb{E}\mu[D(X)] - \mathbb{E}\nu[D(g(Z))]. \tag{2}$$

On peut voir la classe $\mathcal{D}$ comme un sous-ensemble d'une classe plus large $\mathcal{F}$, avec $d_{\mathcal{F}}$ comme une métrique sur les distributions. Divers types de classes $\mathcal{F}$ ont été utilisés dans la littérature sur les GAN. Cela inclut les fonctions continues de Lipschitz (WGAN, (Arjovsky et al., 2017)), les fonctions de Sobolev (Sobolev GAN, (Mroueh et al., 2017)) et l'espace de Hilbert à noyau reproducteur (MMD GAN, (Li et al., 2017)). Ces différences sont à contraster avec celles plus historiquement utilisées dans l'estimation de densité non paramétrique classique, telles que la distance $L^p$, la distance de Hellinger et la divergence de Kullback-Leibler (Tsybakov, 2004), qui ne sont applicables que sous un modèle de domination. Dans le présent article, nous travaillons dans un cadre général où la mesure cible $\mu$ peut avoir une structure de basse dimension rendant ces mesures de divergence habituelles non pertinence. Nous choisissons la classe discriminative $\mathcal{F}$ comme étant la classe Hölder $\mathcal{H}_1^\gamma(\mathbb{R}^p, \mathbb{R})$ correspondant à la boule unité de fonctions de régularité $\gamma \geq 1$. On remarque que le cas $\gamma = 1$ est équivalent au cas de Wasserstein lorsque $\mu$ a un support compact.

On sait que les taux optimaux d'estimation d'une mesure $\mu$ avec un support dans $\mathbb{R}^p$ décroissent de manière exponentielle à mesure que la dimension ambiante $p$ augmente. Pour surmonter cette "curse of dimensionality", certaines hypothèses structurelles de basse dimension sur $\mu$ doivent être imposées. Dans ce travail, nous supposons qu'il existe une application $g^\star \in \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^p)$ avec $\mathbb{T}^d$ le tore $d$-dimensionnel, telle que $\mu = g^\star_{\#U}$ avec $U \sim \mathcal{U}([0,1]^d)$ une variable aléatoire uniforme sur le cube. En particulier, il existe $Y_1, ..., Y_n$ i.i.d. tels que $Y_i \sim \mathcal{U}([0,1]^d)$ et $g^\star(Y_i) = X_i$. Notons que les $Y_i$ sont inconnus, nous n'avons accès qu'aux $X_i$. Dans ce contexte, le problème d'inférence consiste à essayer de trouver un estimateur $\hat{\mu}$ de $\mu = g^\star_{\#U}$ basé sur l'échantillon $(X_i)i = 1, ..., n$, de telle sorte que l'erreur attendue

$$\mathbb{E}_{X_i \sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{\mu}(X_1, ..., X_n))],$$

soit aussi petite que possible. L'estimateur $\hat{\mu}$ est dit minimax optimal s'il n'existe aucun estimateur qui atteint un meilleur taux de convergence uniforme sur le modèle. Formellement, cela signifie qu'il existe une constante $C > 0$ indépendante de $n$ telle que

$$\begin{aligned}&\sup_{g^\star \in \mathcal{H}_K^{\beta+1}} \mathbb{E}_{X_i \sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{\mu}(X_1, ..., X_n))] \\ &\leq C \inf_{\hat{T}} \sup_{g^\star \in \mathcal{H}_K^{\beta+1}} \mathbb{E}_{X_i \sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{T}(X_1, ..., X_n))].\end{aligned}$$

Dans Divol (2021), l'auteur fournit un estimateur minimax (non génératif) des densités dans le cadre de la variété pour la distance de Wasserstein. Il utilise un estimateur local polynomial de la variété provenant de (Aamari and Levrard, 2017) couplé avec un estimateur de densité à noyau. Cependant, l'estimateur polynomial local est très coûteux en termes de calcul et ne peut donc pas être utilisé en haute dimension. Dans Tang and Yang (2022), l'auteur fournit également un estimateur minimax des densités dans le cadre de la variété mais pour la distance $d_{\mathcal{H}_1^\gamma}$. L'estimateur utilise une régularisation de la mesure empirique $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ avec un estimateur de variété couplé à une régularisation des ondelettes tronquée. Cet estimateur est purement théorique et serait extrêmement coûteux à mettre en œuvre en pratique. Par conséquent, l'existence d'un estimateur minimax calculable est toujours une question ouverte et cruciale à résoudre pour fournir des outils efficaces pour des applications concrètes.

Comme l'approche Wasserstein GAN (Arjovsky et al., 2017) s'est avérée facilement implémentable et a fourni des résultats de pointe dans divers domaines, nous nous concentrons dans cet article sur l'estimateur GAN

$$\hat{g} \in \arg\min_{g\in\mathcal{G}} \sup_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} D(X_i) - D(g(U_i)) \tag{3}$$

pour $U_i \sim \mathcal{U}([0,1]^d)$ i.i.d., $\mathcal{G} \subset \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^p)$ et $\mathcal{D} \subset \mathcal{H}_1^{\gamma}(\mathbb{R}^p, \mathbb{R})$ avec $\gamma \in [1, \beta+1]$. L'estimateur $\hat{g}$ doit être compris comme une approximation empirique de la solution de (2) basée sur les données $X_1, ..., X_n$. Les mesures de probabilité $\hat{g}_{\#U}$ sont alors naturellement nos estimateurs de la cible $g^\star_{\#U}$. Rappelons qu'étant donné que $\mathcal{D} \subset \mathcal{H}^\gamma$, le cas $\gamma = 1$ correspond à l'estimateur classique WGAN. L'une des forces de l'estimateur GAN (3) est qu'il effectue à la fois des estimations de support et de densité en même temps, ce qui permet en particulier d'éviter l'utilisation de tout estimateur de variété comme dans Divol (2021) et Tang and Yang (2022).

## Contributions principales

Nous construisons des classes de réseaux neuronaux calculable $\mathcal{G}$ et $\mathcal{D}$, telles que

$$\sup_{g^\star\in\mathcal{H}_K^{\beta+1}} \mathbb{E}_{X_i\sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{g}_{\#U})] \leq C_1(\log n)^{C_2}\left(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}}\right), \tag{4}$$

avec $C_1, C_2 > 0$ des constantes indépendantes de $n$. Ce taux a été prouvé comme étant minimax optimal dans Tang and Yang (2022) jusqu'aux facteurs logarithmiques. À notre connaissance, il s'agit de la première étude montrant que l'estimateur GAN atteint les taux minimax pour les distances Wasserstein/Hölder. Ce résultat améliore, en particulier, les taux obtenus dans Chen et al. (2020), Schreuder et al. (2021) et Chae (2022). Les taux de convergence minimax de la version classique du GAN (Goodfellow et al., 2014) ont récemment été obtenus par Belomestny et al. (2023) pour la divergence Jensen–Shannon. Leur résultat traite uniquement du cadre de dimension pleine $d = p$, car la divergence Jensen–Shannon n'est non triviale que lorsque les mesures à comparer ne sont pas singulières l'une par rapport à l'autre.

Nous obtenons le taux de convergence minimax (4) sur les mesures $\mu = g^\star_{\#U}$ qui ont une densité par rapport à la mesure du volume d'une sous-variété inconnue. Dans le processus, des taux minimax sont également obtenus pour deux modèles statistiques intermédiaires :

- Nous traitons d'abord un cadre d basse dimension où la mesure cible peut avoir des atomes et son support n'est pas nécessairement une variété. Dans ce cas, la classe des discriminateurs $\mathcal{D}$ utilisée est théorique, ce qui signifie qu'elle n'est pas calculable en pratique. Ce modèle est étudié pour discuter des limites des hypothèses du cas général.
- Nous prouvons des taux minimax dans le cadre de la dimension pleine $p = d$, où la mesure cible a une densité par rapport à la mesure de Lebesgue $p$-dimensionnelle. Ce cas est traité pour comprendre dans le cadre plus simple de la dimension pleine comment des hypothèses supplémentaires peuvent nous aider à obtenir un estimateur calculabe.

- En adaptant la méthode développée dans le cas de dimension pleine au cas de la sous-variété, nous proposons un estimateur GAN calculable atteignant des taux minimax pour tous les $\gamma \in [1, \beta + 1]$ simultanément.

Le résultat principal est démontré à l'aide d'une nouvelle inégalité d'interpolation qui borne la distance $d_{\mathcal{H}_1^\gamma}(g_{\#U}, g^\star_{\#U})$ par la quantité $d_{\mathcal{H}_1^{\beta+1}}(g_{\#U}, g^\star_{\#U})^{\frac{\beta+\gamma}{2\beta+1}}$ et un facteur logarithmique.

# References


Eddie Aamari and Clément Levrard. Non-asymptotic rates for manifold, tangent space, and curvature estimation, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y.W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR, 2017.

Denis Belomestny, Eric Moulines, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Rates of convergence for density estimation with generative adversarial networks, 2023.

Minwoo Chae. Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. *arXiv preprint arXiv:2202.02890*, 2022.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.

Vincent Divol. Measure estimation on manifolds: an optimal transport approach, 2021.

I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A.C. Courville. Improved training of Wasserstein GANs. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777. Curran Associates, Inc., 2017.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *CoRR*, abs/2106.12423, 2021.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network, 2017.

Yufei Liu, Yuan Zhou, Xin Liu, Fang Dong, Chang Wang, and Zihong Wang. Wasserstein gan-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering*, 5(1):156–163, 2019.

Yun Luo and Bao-Liang Lu. Eeg data augmentation for emotion recognition using a conditional wasserstein gan. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2535–2538. IEEE, 2018.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan, 2017.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pages 1051–1071. PMLR, 2021.

Jan Stanczuk, Christian Etmann, Lisa Maria Kreusser, and Carola-Bibiane Schönlieb. Wasserstein gans work because they fail (to approximate the wasserstein distance). *arXiv preprint arXiv:2103.01678*, 2021.

Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifold under adversarial losses, 2022.

Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 613–621. Curran Associates, Inc., 2016.

L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2852–2858. AAAI Press, 2017.

# Analyse de la force de bruitage dans les modèles génératifs basés sur le score.

Stanislas Strasman*, Antonio Ocello†, Claire Boyer*,‡, Sylvain Le Corff* & Vincent Lemaire*

* *LPSM, Sorbonne Université, UMR CNRS 8001, France,*
*stanislas.strasman@sorbonne-universite.fr*
*prenom.nom@sorbonne-universite.fr*
† *CMAP, École Polytechnique, Institut Polytechnique de Paris, France,*
*prenom.nom@polytechnique.edu*
‡ *Institut Universitaire de France (IUF)*

**Résumé.** Les modèles génératifs basés sur le score (SGMs) visent à estimer une distribution de données cible en apprenant des fonctions de score (correspondant au gradient du logarithme de densités de probabilité) uniquement à partir d'échantillons bruités de la cible. La littérature récente s'est largement concentrée sur l'évaluation de l'erreur entre les distributions cible et estimée, en mesurant la qualité des données générées à travers la divergence de Kullback-Leibler (KL) ou encore des distances de Wasserstein. Néanmoins, les résultats existants ont été obtenus pour une vitesse de bruitage homogène dans le temps. Sous des hypothèses faibles sur la distribution des données, nous établissons une borne supérieure pour la divergence KL entre les distributions cible et estimée qui dépend explicitement de la force de bruitage utilisée au cours du temps. De plus, en supposant que le score est Lipschitz continu, et avec des capacités d'approximation du score et de discrétisation parfaites, nous montrons une borne d'erreur améliorée en distance de Wasserstein, tirant parti des mécanismes de contraction sous-jacents des équations différentielles stochastiques en jeu. Enfin, nous proposons un algorithme pour ajuster automatiquement la fonction de bruitage au cours de la diffusion en utilisant la borne supérieure théorique établie.

**Mots-clés.** Méthodes générative diffusives, modèles génératifs basés sur le score, force de bruitage.

**Abstract.** Score-based generative models (SGMs) aim at estimating a target data distribution by learning score functions using only noise-perturbed samples from the target. Recent literature has focused extensively on assessing the error between the target and estimated distributions, gauging the generative quality through the Kullback-Leibler (KL) divergence and Wasserstein distances. All existing results have been obtained so far for time-homogeneous speed of the noise schedule. Under mild assumptions on the data distribution, we establish an upper bound for the KL divergence between the target and the estimated distributions, explicitly depending on any time-dependent noise schedule. Assuming that the score is Lipschitz continuous, we provide an improved error bound in Wasserstein distance, taking advantage of favourable underlying contraction mechanisms. We also propose an algorithm to automatically tune the noise schedule using the proposed upper bound. We illustrate empirically the performance of the noise schedule optimization in comparison to standard choices in the literature.

**Keywords.** Generative diffusion models, score-based generative models, noise schedule.

# 1 Introduction

Les modèles génératifs visent à générer de nouveaux échantillons synthétiques à partir d'échantillons dits d'entraînement, supposés être issus d'une distribution $\pi_{\text{data}}$ inconnue. Ces modèles incluent des approches reposant sur des diffusions (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021), qui ont récemment permis des résultats prometteurs dans de nombreuses applications, comme par exemple dans le cadre de la génération d'image à partir de description textuelle (Ramesh et al., 2022), ou encore de la génération de langage naturel (Gong et al., 2023). Nous référons le lecteur à Yang et al. (2023) qui offre un aperçu complet des dernières avancées sur ce sujet. Il convient de noter que dans ces applications réelles, la complexité des données empêche en général de représenter la distribution $\pi_{\text{data}}$ à travers un modèle paramétrique classique et, proscrit dès lors toute estimation par des méthodes de maximum de vraisemblance traditionnelles. Les modèles génératifs basés sur le score (SGMs) représentent alors une alternative implémentable.

**Modèles génératifs utilisant le score (SGM).** Les modèles génératifs basés sur le score (SGMs) sont des modèles probabilistes s'articulant en deux phases. La première phase consiste à bruiter les données (également appelée phase *forward*), c'est-à-dire que l'on perturbe progressivement la distribution empirique en ajoutant du bruit aux données d'entraînement jusqu'à ce que leur distribution atteigne approximativement une distribution facilement échantillonnable $\pi_\infty$. D'autre part, la seconde phase vise à inverser cette dynamique en débruitant séquentiellement des réalisations de $\pi_\infty$. On appelle cette phase, la phase d'échantillonnage (ou phase *backward*). Inverser la dynamique nécessite en principe la connaissance de la fonction de score, c'est-à-dire, le gradient du logarithme de la densité du processus *forward* à chaque étape temporelle de la diffusion. Cependant, connaître le score exactement revient à connaître la distribution au temps $t = 0$, c'est-à-dire, à connaître la distribution cible $\pi_{\text{data}}$ selon laquelle nous souhaitons simuler de nouveaux exemples. Pour contourner ce problème, la fonction de score est donc apprise sur la base de l'évolution des échantillons de données bruités, en utilisant une architecture de réseau de neurones profond. Une fois le score appris, nous pouvons l'utiliser dans la dynamique inverse appliquée aux échantillons tirés selon $\pi_\infty$, nous obtenons ainsi un modèle génératif, approchant des tirages selon $\pi_{\text{data}}$.

Une attention significative a été accordée à la compréhension des sources d'erreurs qui affectent la qualité de la génération de données associée aux SGM (Chen et al., 2023a,b; De Bortoli, 2022; Lee et al., 2023). Pour ce faire, des bornes supérieures pour des (pseudo-)distances entre les distributions d'échantillons d'entraînement et générés ont été établies.

**Contributions.** Dans ce travail, nous procédons à une analyse mathématique approfondie de la force de bruitage dans les modèles génératifs basés sur le score.

— Nous établissons une borne supérieure pour la divergence de Kullback-Leibler (KL) entre la distribution des données et la loi du SGM. Cette borne est valide sous des hypothèses minimales et dépend explicitement de la force de bruitage utilisée pour entraîner le SGM.
— À travers des expériences numériques, nous illustrons la borne supérieure obtenue en pratique en ce qui concerne les divergences KL empiriques effectives. Ces simulations mettent en évidence la pertinence de la borne supérieure, reflétant en pratique l'effet de la force de bruitage au cours du temps sur la qualité de la distribution générée.
— En faisant une hypothèse supplémentaire sur la régularité de la fonction de score, nous établissons une borne plus précise de l'erreur due au temps de mélange en termes de distance de Wasserstein, en tirant parti de la contraction du terme de dérive non seulement en phase forward, mais aussi en en phase backward de la diffusion stochastique.
— Enfin, nous proposons d'exploiter la borne théorique obtenue pour guider et améliorer la mise en œuvre des SGM en pratique. Nous suggérons en effet une procédure pour optimiser conjointement le réseau de neurones approchant le score et la force de bruitage en utilisant comme fonction objectif la borne supérieure établie.

# 2 Analyse théorique de l'impact de la force de bruitage dans les SGMs

## 2.1 Notation et définitions

**Processus *forward*.** Posons $\beta : [0, T] \mapsto \mathbb{R}_{>0}$ la fonction de bruitage, supposée continue et croissante. Bien que développés initialement en utilisant un nombre fini d'étapes de bruitage, les analyses les plus récentes considèrent des perturbations continues à l'aide d'équations différentielles stochastiques (EDS) (Song et al., 2021). Considérons donc un processus *forward* donné par

$$\mathrm{d}\overrightarrow{X}_t = -\frac{\beta(t)}{2\sigma^2}\overrightarrow{X}_t \mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_{\text{data}}\,. \tag{1}$$

Notons $p_t$ la densité de $\overrightarrow{X}_t$ au temps $t \in (0, T]$. Comme la dérive est linéaire par rapport à $(X_t)_{t\geq 0}$, une simulation exacte de ce processus est possible. De plus, la distribution stationnaire du processus *forward* est la distribution gaussienne avec moyenne 0 et variance $\sigma^2 \mathrm{I}_d$ que l'on note $\pi_\infty$.

Lorsque $\beta(t)$ est constant, égal à 2, (c'est-à-dire dans le cas homogène en temps), ce processus de diffusion est connu sous le nom de *Variance Preserving SDE* (VPSDE, De Bortoli et al., 2021; Conforti et al., 2023; Chen et al., 2023b), dont la discrétisation permet de retrouver les *Denoising Diffusion Probabilistic Models* (DDPM, Ho et al., 2020).

**Processus *backward*.** Le processus *backward* correspondant est initialisé à la distribution stationnaire $\pi_\infty$ et peut être écrit

$$\mathrm{d}\overleftarrow{X}_t = \eta(t, \overleftarrow{X}t)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim \pi_\infty,$$

où

$$\begin{cases} \bar{\beta}(t) & := \beta(T-t) \\ \eta(t, \overleftarrow{X}_t) & := \bar{\beta}(t)\overleftarrow{X}_t/(2\sigma^2) + \bar{\beta}(t)\nabla \log p_{T-t}\left(\overleftarrow{X}_t\right). \end{cases}$$

Notons $\mathbb{Q}_T \in \mathcal{P}(C([0,T], \mathbb{R}^d))$ la mesure de chemin associée à la diffusion *backward*. Nous introduisons $\tilde{p}_t(x)$ la distribution marginale (en temps) du processus *forward* renormalisée par la densité de sa distribution stationnaire :

$$\forall x \in \mathbb{R}^d, \quad \tilde{p}_t(x) = \frac{p_t(x)}{\varphi_{\sigma^2}(x)}, \tag{2}$$

où $\varphi_{\sigma^2}$ désigne la fonction de densité de $\pi_\infty$, une distribution gaussienne avec moyenne 0 et variance $\sigma^2 \mathrm{I}_d$. Ainsi, le processus *backward* peut être réécrit

$$\mathrm{d}\overleftarrow{X}_t = \bar{\eta}\left(t, \overleftarrow{X}t\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim \pi_\infty, \tag{3}$$

où $\bar{\eta}(t, \overleftarrow{X}_t) := -\frac{\bar{\beta}(t)}{2\sigma^2}\overleftarrow{X}_t + \bar{\beta}(t)\nabla \log \tilde{p}_{T-t}(\overleftarrow{X}_t)$. Considérer $\tilde{p}_t$ dans notre analyse permet de ré-interpréter le processus *backward* comme une perturbation d'un processus OU. Cette astuce est cruciale pour mettre en valeur le rôle central de l'information de Fisher dans la performance du SGM. Cette renormalisation a déjà été utilisée par Conforti et al. (2023).

**Estimation du score.** Simuler le processus *backward* requiert de connaître le score. Cependant, la fonction de score (modifiée) $\nabla \log \tilde{p}_t(x) = \nabla \log p_t(x) + x/\sigma^2$ ne peut pas être évaluée directement, car elle dépend de la distribution des données inconnues. Pour contourner ce problème, la fonction de score $\nabla \log p_t$ doit être estimée. Dans Hyvärinen and Dayan (2005), les auteurs proposent d'estimer la fonction de score associée à une distribution en minimisant la distance au carré $\mathrm{L}^2$ entre la vraie fonction de score et l'approximation proposée. Dans le contexte des modèles de diffusion, cela se fait généralement avec l'utilisation d'une architecture de réseau de neurones profond $s_\theta : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ paramétrée par $\theta \in \Theta$, et visant à minimiser :

$$\mathcal{L}_{\text{explicit}}(\theta) = \mathbb{E}\left[\left\| s_\theta(\tau, \overrightarrow{X}_\tau) - \nabla \log p_\tau(\overrightarrow{X}_\tau)\right\|^2\right],$$

avec $\tau \sim \mathcal{U}(0,T)$ indépendant du processus *forward* $(\overrightarrow{X}_t)_{t\geq 0}$. Cependant, ce problème d'estimation souffre du fait que la cible de régression n'est pas explicitement connue. Un problème d'optimisation tractable partageant les mêmes optima peut cependant être défini à travers la marginalisation de $\pi_{\text{data}}$ par $p_\tau$ (voir Vincent, 2011; Song et al., 2021) :

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E}\left[\left\| s_\theta(\tau, \overrightarrow{X}_\tau) - \nabla \log p_\tau(\overrightarrow{X}_\tau | X_0)\right\|^2\right],$$

où $\tau$ est uniformément distribué sur $[0,T]$, indépendant de $X_0 \sim \pi_{\text{data}}$ et $\overrightarrow{X}\tau \sim p_\tau(\cdot|X_0)$. Cette fonction de coût ne requiert que la connaissance du noyau de transition du processus *forward*. Dans le cadre classique des modèles de diffusion donné par (1), il s'agit d'un noyau gaussien avec moyenne et variance explicites.

**Discrétisation.** Une fois la fonction de score apprise, il reste que la dynamique *backward* ne présente plus une dérive linéaire. Cela rend sa simulation exacte difficile. Pour résoudre ce problème, une solution consiste à discrétiser la dynamique continue du processus *backward*. Ainsi, Song et al. (2021) propose un schéma de discrétisation d'Euler-Maruyama (EM) dans lequel les coefficients de dérive et de diffusion sont discrétisés récursivement. En particulier, introduisons $\tilde{s}_\theta(t,x) := s_\theta(t,x) + x/\sigma^2$ et considérons la discrétisation temporelle $0 =: t_0 \leq t_1 \leq \cdots \leq t_N := T$, le schéma EM correspond à

$$\mathrm{d}\overleftarrow{X}_t^{EM} = \left(-\frac{\bar{\beta}(t_k)}{2\sigma^2}\overleftarrow{X}_{t_k}^{EM} + \bar{\beta}(t)\tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}^{EM}\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t_k)}\mathrm{d}B_t\,.$$

Autrement, l'intégrateur exponentiel d'Euler (EI), déjà utilisé dans Conforti et al. (2023), nécessite seulement de discrétiser la partie associée à la fonction de score modifiée. Soit $\left(\overleftarrow{X}_t^\theta\right) t \in [0,T]$ tel que, pour $t \in [t_k, t_{k+1}]$,

$$\mathrm{d}\overleftarrow{X}_t^\theta = \bar{\beta}(t)\left(-\frac{1}{2\sigma^2}\overleftarrow{X}_t^\theta + \tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}^\theta\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t\,.$$

Ce schéma peut être considéré comme un raffinement du schéma classique Euler-Maruyama car il intègre explicitement le terme de dérive linéaire. Nous considérons donc un tel schéma dans nos développements théoriques ultérieurs. Enfin, notons $\bar{\mathbb{Q}}_N^{\beta,\theta} \in \mathcal{P}(C([0,T],\mathbb{R}^d))$ la mesure de chemin associée à cette version discrétisée de la diffusion *backward* et par $\widehat{\pi}_N^{(\beta,\theta)}$ la densité de probabilité marginale de $\overleftarrow{X}_T^\theta$ avec une discrétisation à $N$ pas de temps.

## 2.2 Une borne supérieure explicite en la force de bruitage

La distribution des données $\pi_{\text{data}}$ est supposée absolument continue par rapport à la mesure gaussienne $\pi_\infty$. L'information de Fisher relative $\mathcal{I}(\pi_{\text{data}}|\pi_\infty)$ est donnée par

$$\mathcal{I}(\pi_{\text{data}}|\pi_\infty) := \int \left\|\nabla \log\left(\frac{\mathrm{d}\pi_{\text{data}}}{\mathrm{d}\pi_\infty}\right)\right\|^2 \mathrm{d}\pi_{\text{data}}\,.$$

Nous considérons les hypothèses suivantes :

**H1** La fonction de bruitage est continue, non décroissante et telle que $\int_0^\infty \beta(t)\mathrm{d}t = \infty$.

**H2** La distribution des données a une information de Fisher finie par rapport à la distribution gaussien, c'est-à-dire, $\mathcal{I}(\pi_{\text{data}}|\pi_\infty) < \infty$.

**H3** Le paramètre $\theta \in \Theta$ et la fonction $\beta$ satisfont

$$\mathbb{E}\left[\exp\left\{\frac{1}{2}\int_0^T \bar{\beta}(t)\left\|\left(\tilde{s}\left(T-t, \overleftarrow{X}_t\right) - \tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}\right)\right)\right\|^2 \mathrm{d}t\right\}\right] < \infty,$$

où $\tilde{s}(t,x) := \nabla \log \tilde{p}_t(x)$ correspond à la fonction de score modifiée (2).

L'hypothèse H1 est nécessaire pour garantir que le processus *forward* converge vers la distribution stationnaire lorsque le temps de diffusion tend vers l'infini. L'hypothèse H2 est inhérente à la distribution des données, car elle implique seulement l'intégrabilité $L^2$ de la fonction de score. Un tel type d'hypothèse a déjà été considéré dans la littérature, voir Conforti et al. (2023). Enfin, l'hypothèse H3 garantit une bonne approximation du score par le réseau de neurones $\tilde{s}_\theta$, pondérée par le niveau de bruitage.

**Theorem 2.1.** *Supposons que H1, H2 et H3 soient vérifiées. Alors,*

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) \leq \mathcal{E}_1(\beta) + \mathcal{E}_2(\theta,\beta) + \mathcal{E}_3(\beta)\,,$$

*où*

$$\begin{aligned}
\mathcal{E}_1(\beta) &= \mathrm{KL}\left(\pi_{\mathrm{data}} || \pi_\infty\right) \exp\left\{-\frac{1}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right\}\,,\\
\mathcal{E}_2(\theta,\beta) &= \sum_{k=1}^{N} \mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t_k}\left(\overrightarrow{X}_{T-t_k}\right) - \tilde{s}_\theta\left(T-t_k, \overrightarrow{X}_{T-t_k}\right)\right\|^2\right] \int_{t_k}^{t_{k+1}} \beta(t)\mathrm{d}t\,,\\
\mathcal{E}_3(\beta) &= 2h\beta(T)\max\left\{\frac{h\beta(T)}{4\sigma^2}; 1\right\} \mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty)\,,
\end{aligned}$$

*avec* $h := \sup_{k\in 1,\dots,N}(t_k - t_{k-1})$ *et* $t_0 := 0$.

La borne obtenue permet de retrouver des garanties existantes (De Bortoli et al., 2021; Conforti et al., 2023), mais va au-delà en faisant apparaître une dépendance explicite en la force de bruitage $\beta$, soit à travers sa version intégrée sur le temps de diffusion, soit à travers sa valeur finale au temps $T$.

Les différents termes de la borne correspondent à trois types d'erreurs affectant les performances des SGMs. Le terme $\mathcal{E}_1$ représente le *temps de mélange* du processus forward d'Ornstein-Uhlenbeck, résultant de la limitation pratique de considérer le processus forward jusqu'à un temps fini $T$. En effet, $\mathcal{E}_1$ tend vers 0 lorsque $T$ tend vers l'infini. Notons que le terme multiplicatif dans $\mathcal{E}_1$ correspond à la divergence KL entre $\pi_{\mathrm{data}}$ et $\pi_\infty$, finie par l'Hypothèse H2. Le second terme $\mathcal{E}_2$ correspond à l'*erreur d'approximation*, qui découle de l'utilisation d'un réseau de neurones pour estimer la fonction de score. Enfin, $\mathcal{E}_3$ est l'*erreur de discrétisation* du schéma de discrétisation EI. Ce dernier terme disparaît à mesure que la grille de discrétisation est de plus en plus fine (c'est-à-dire, $h \to 0$).

# 3 Sur la finesse de la borne supérieure

## 3.1 Une version raffinée

Dans cette section, nous nous concentrons sur le cadre d'une "approximation parfaite du score" et d'une discrétisation infiniment précise, c'est-à-dire, $\mathcal{E}_2(\theta,\beta) = \mathcal{E}_3(\theta,\beta) = 0$. Cela permet d'évaluer la précision du terme $\mathcal{E}_1(\beta)$ dans la borne supérieure du Théorème 2.1.

Lorsque la distribution des données est restreinte à une gaussienne $\mathcal{N}(\mu_0, \Sigma_0)$, on peut exploiter la contraction *backward* en supposant que $\lambda_{\max}(\Sigma_0) \leq \sigma^2$, où $\lambda_{\max}(\Sigma_0)$ désigne la plus grande valeur propre de $\Sigma_0$. Dans ce cas spécifique, nous pouvons obtenir une version raffinée pour $\mathcal{E}_1$, donnée par

$$\mathrm{KL}\left(\pi\mathrm{data}|\varphi_{\sigma^2}Q_T\right) \leq \mathrm{KL}\left(\pi_{\mathrm{data}}|\varphi_{\sigma^2}\right)\exp\left(-\frac{2}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right).$$

En outre, dans la littérature, d'autres bornes supérieures ont été obtenues pour d'autres métriques telles que les distances de Wasserstein (Lee et al., 2023; De Bortoli, 2022). Nous proposons un contrôle en distance de Wasserstein sous l'hypothèse suivante :

**H4** For any $t$, there exists $C_t \geq 0$ such that $\forall x, y \in \mathbb{R}^d$,

$$\left(\nabla \log \tilde{p}_t(x) - \nabla \log \tilde{p}_t(y)\right)^\top (x-y) \leq -C_t \left\|x-y\right\|^2 .$$

**Proposition 3.1.** *Supposons que $x \mapsto \nabla \log \tilde{p}_t(x)$ soit Lipschitz continu de coefficient $L_t$ pour $t \in (0, T]$. Alors,*

$$\mathcal{W}_2(\pi_{\mathrm{data}}, \varphi_{\sigma^2}Q_T)^2 \leq \mathcal{W}_2(p_T, \varphi_{\sigma^2})^2 \times \exp\left(-\int_0^T \frac{\bar{\beta}(t)}{\sigma^2}\left(1 - 2L_t\sigma^2\right)\mathrm{d}t\right). \tag{4}$$

*De plus, sous l'Hypothèse H4, nous avons*

$$\mathcal{W}_2(\pi_{\mathrm{data}}, \varphi_{\sigma^2}Q_T)^2 \leq \mathcal{W}_2(p_T, \varphi_{\sigma^2})^2 \times \exp\left(-\int_0^T \frac{\bar{\beta}(t)}{\sigma^2}\left(1 + 2C_t\sigma^2\right)\mathrm{d}t\right). \tag{5}$$

Notons que l'Hypothèse H4 ou la propriété Lipschitz de la fonction de score sont toutes deux satisfaites lorsque la distribution cible est supposée être gaussienne avec une structure de covariance appropriée.

**Lemma 3.2.** *Supposons que $\pi_{\mathrm{data}}$ soit une distribution gaussienne $\mathcal{N}(\mu_0, \Sigma_0)$, telle que $\lambda_{\max}(\Sigma_0) \leq \sigma^2$. Alors, la borne d'erreur (5) est valable avec une contraction donnée par la constante suivante*

$$C_t := \frac{m_t^2\left(\sigma^2 - \lambda_{\max}(\Sigma_0)\right)}{m_t^2\lambda_{\max}(\Sigma_0) + \sigma^2\left(1 - m_t^2\right)}.$$

Ce résultat, restreint au cas gaussien, met l'accent sur l'importance de calibrer le paramètre $\sigma^2$ en fonction de la structure de covariance de la distribution des données, afin d'accélérer la vitesse de convergence de l'algorithme.

## 3.2 Illustration numérique

Pour illustrer la borne supérieure, nous considérons le cas où la vraie distribution est gaussienne en dimension $d = 50$ avec une moyenne $\mathbf{1}_d$ et différents choix de structure de covariance :

1. (Isotrope) $\Sigma^{(\mathrm{iso})} = 0.5\mathrm{I}_d$.
2. (Hétéroscédastique) $\Sigma^{(\mathrm{heterosc})} \in \mathbb{R}^{d\times d}$ est une matrice diagonale telle que $\Sigma^{(\mathrm{heterosc})}_{jj} = 10$ pour $1 \le j \le 5$, et $\Sigma^{(\mathrm{heterosc})}_{jj} = 0.1$ sinon.
3. (Corrélation) $\Sigma^{(\mathrm{corr})} \in \mathbb{R}^{d\times d}$ est une matrice pleine dont les entrées diagonales sont égales à un et les termes hors diagonale sont $\Sigma^{(\mathrm{corr})}_{jj'} = 1/\sqrt{|j-j'|}$ pour $1 \le j \ne j' \le d$.

Les distributions de données résultantes sont respectivement désignées par $\pi^{(\mathrm{iso})}_{\mathrm{data}}$, $\pi^{(\mathrm{heterosc})}_{\mathrm{data}}$ et $\pi^{(\mathrm{corr})}_{\mathrm{data}}$. Le Théorème 2.1 fournit une borne supérieure générique de Kullback-Leibler :

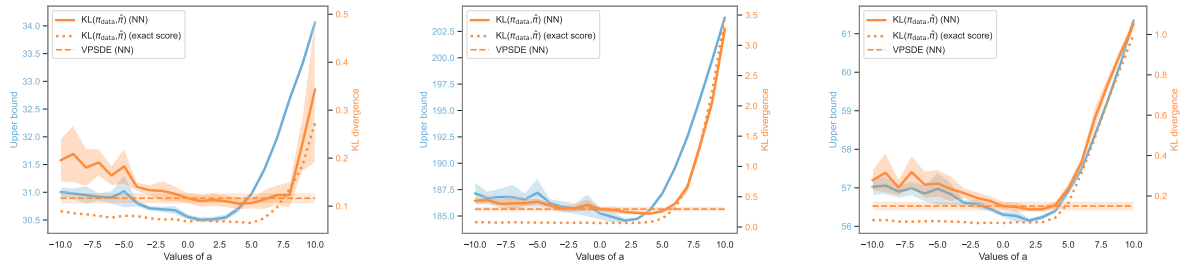
$$\mathcal{L}_{\mathrm{sched}}(\theta,\beta) = \mathcal{E}_1(\beta) + \mathcal{E}_2(\theta,\beta) + \mathcal{E}_3(\beta). \tag{6}$$

Nous proposons d'évaluer (6) pour les différentes distributions de données ci-dessus, et pour une fonction de bruitage de la forme

$$\beta_a(t) \propto (\mathrm{e}^{at} - 1)/(\mathrm{e}^{aT} - 1), \tag{7}$$

avec $a \in \mathbb{R}$ variant de $-10$ à 10. Pour ce faire, pour chaque valeur de $a$, et chaque distribution de données, nous entraînons un SGM avec 200 étapes de discrétisation de l'intervalle de temps $[0,1]$ avec $n = 10000$ échantillons gaussiens. Le score est appris en utilisant un réseau de neurones dense avec 3 couches cachées de largeur 256 sur 100 époques.

La Figure 1 met en évidence dans tous les scénarios que la force de bruitage utilisée dans les SGM impacte la valeur de $\mathrm{KL}(\pi_{\mathrm{data}} \| \widehat{\pi}_N^{(\beta,\theta)})$, et donc la qualité de la distribution apprise.




(a) Cas isotrope (b) Cas hétéroscédastique (c) Cas corrélé

FIGURE 1 – Comparaison de la divergence KL empirique (valeur moyenne ± écart-type sur 10 réalisations) entre $\pi_{\mathrm{data}}$ et $\widehat{\pi}_N^{(\beta,\theta)}$ (en orange) et la borne supérieure (6) (en bleu) par rapport au paramètre $a$ utilisé dans la définition de la force de bruitage $\beta_a$, pour $d = 50$. Nous représentons également la divergence KL obtenue avec le modèle *VPSDE* (ligne *dashed*) et celle obtenue avec notre modèle (ligne *dotted*) lorsque le score n'est pas approché mais directement évalué.

# 4 Optimisation de la force de bruitage

**Algorithme.** Nous proposons d'exploiter la borne supérieure théorique (6) pour ajuster le choix de la force de bruitage. À cette fin, nous proposons une méthode itérative pour optimiser

conjointement les poids $\theta$ du réseau de neurones et la force de bruitage $\beta$, voir Algorithme 1. Les fonctions admissibles $\beta_a$ pour la fonction de bruitage sont données par (7). Pour des comparaisons justes, nous entraînons à la fois le réseau *VPSDE* et le réseau adaptatif avec 10000 échantillons sur 200 époques en utilisant le même *learning rate*.

**Algorithme 1** Optimisation itérative de la force de bruitage et de la fonction de score

**Entrée :** $N$ échantillons d'entraînement, programme initial $\beta_a$ avec $a = a^{(0)}$, paramètre initial $\theta^{(0)}$.
Définir $a^\star = a^{(0)}$
**for** $e = 0$ **to** nombre d'époques **do**
  Calculer $\theta^{(e+1)}$ en utilisant un *score matching* avec la fonction de bruitage $\beta_{a^\star}$ et l'estimation initiale $\theta^{(e)}$.
  **if** $e$ mod10 $= 0$ **then**
    Mettre à jour

$$a^\star \in \operatorname{argmin}_a \mathcal{L}_{\text{sched}}(\theta^{(e+1)}, \beta_a).$$

  **end if**
**end for**

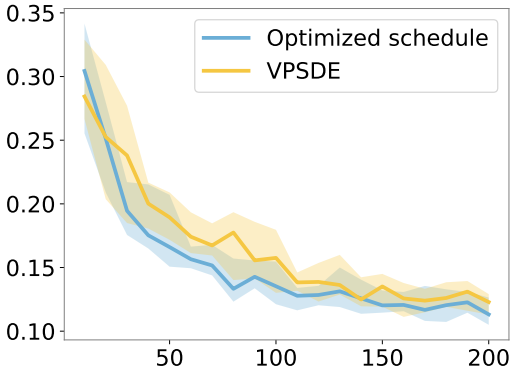



FIGURE 2 – Divergences KL empiriques (médiane et quartiles sur 10 exécutions) entre $\pi_{\text{data}}$ et les distributions obtenues par l'Algorithme 1 (bleu) et le modèle VPSDE (jaune).

**Résultats.** Nous évaluons la performance de l'Algorithme 1 en considérant une distribution cible de données gaussienne $\pi_{\text{data}}^{(\text{corr})}$.

Sur la Figure 2, au fil des époques, nous affichons les divergences KL empiriques par rapport à la distribution générée via l'Algorithme 1 et par rapport à un algorithme *VPSDE* classique. Dès les premières époques, l'Algorithme 1 produit de meilleurs échantillons que le modèle *VPSDE* standard. Comme attendu, la valeur de $a$ sélectionnée par l'Algorithme 1 tend à être décalée vers des valeurs positives avec une certaine stabilisation autour des valeurs optimales déjà observées sur la Figure 1.

# Références

H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling : User-friendly bounds under minimal smoothness assumptions. In International Conference on Machine Learning, pages 4735–4763. PMLR, 2023a.

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score : theory for diffusion models with minimal data assumptions, 2023b.

G. Conforti, A. Durmus, and M. G. Silveri. Score diffusion models without early stopping : finite fisher information is all you need, 2023.

V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. arXiv preprint arXiv :2208.05314, 2022.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. Advances in Neural Information Processing Systems, 34 :17695–17709, 2021.

S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong. Diffuseq : Sequence to sequence text generation with diffusion models. In Proceedings of International Conference on Learning Representations, 2023.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, 2020.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6(4), 2005.

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In International Conference on Algorithmic Learning Theory, pages 946–985. PMLR, 2023.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv :2204.06125, 1(2) :3, 2022.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In Advances in Neural Information Processing Systems, 2019.

Y. W. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.

P. Vincent. A connection between score matching and denoising autoencoders. Neural Computation, 23(7) :1661–1674, 2011. doi : 10.1162/NECO_a_00142.

L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models : A comprehensive survey of methods and applications. ACM Computing Surveys, 56(4) :1–39, 2023.

# Statistique et sport 2

# THÉORIE DES JEUX ET STATISTIQUES SPORTIVES : L'ENSEIGNEMENT DU JEU DU PENALTY

Léo Gerville-Réache[1]

[1] Univ. Bordeaux, IMS, UMR 5218, F-33400, Talence, France
leo.gerville-reache@u-bordeaux.fr

**Résumé.** En partant de statistiques sur les penalties au football (Chiappori et al., 2002), cette communication revisite une proposition de cours introductif portant sur la modélisation des jeux 2x2 à travers la théorie des jeux non coopératifs, simultanés et à sommes constantes. De la construction d'une matrice des gains basée sur les statistiques de matchs à l'analyse mathématique et statistique du jeu, la problématique du penalty semble être un choix pertinent pour introduire les concepts de minimax et d'équilibre de Nash, tout en questionnant la rationalité des joueurs.

**Mots-clés.** Penalty, football, théorie des jeux, minimax, équilibre de Nash.

**Abstract.** Based on statistics on penalties in football (Chiappori et al., 2002), this paper revisits a proposal for an introductory course on modeling 2x2 games through the theory of non-cooperative, simultaneous, constant-sum games. From constructing a payoff matrix based on matches statistics to mathematical and statistical analysis of the game, the penalty dilemma appears to be a relevant choice for introducing the concepts of minimax and Nash equilibrium while questioning the rationality of players.

**Keywords.** Penalty, football, game theory, minimax, Nash equilibrium.

## 1 Introduction

En cherchant quelques instants sur le Web, vous trouverez sûrement quelques statistiques sur la fréquence moyenne de pénaltys par match de football. Il est seulement question de ceux réalisés sur fautes. En fonction des championnats nationaux, elle varie entre 0,23 et 0,45 ; la ligue 1 se situant autour de 0,36. Pour résumer, en ligue 1, on tire en moyenne un pénalty tous les 3 matchs. Si l'on regarde maintenant le pourcentage de buts marqués sur pénalty, il se situe autour de 75%. On peut lire, sur le journal « L'équipe » en ligne publié le 28 février 2019 : « *Sur les 91 000 pénaltys (dans le match) ou tirs au but étudiés, 75,49 % ont été victorieux, 17,57 % ont été repoussés par le gardien, 4,07 % ont raté le cadre, 2,87 % ont été renvoyés par la barre ou un montant*. »

Il n'y a clairement pas de situation de jeu qui donne une statistique aussi favorable à la possibilité de but. Pour autant, seuls 3 sur 4 seront marqués. Cette situation très particulière de jeu permet de s'interroger sur les choix optimaux du gardien et du tireur. Chiappori *et al.* en 2002 publient un article particulièrement intéressant sur l'analyse d'un grand ensemble de pénaltys (de la ligue 1 française et italienne sur 3 année) et son analyse via la théorie des jeux et l'équilibre de Nash. C'est la rationalité des gardiens et des tireurs qui est questionnée. Les auteurs étudient un jeu croisant 2 joueurs (gardien et tireur) et 3 choix : tirer à droite, gauche ou au centre (pour le tireur) et partir à droite, gauche ou au centre (pour le gardien).

Dans une approche pédagogique destinée à des étudiants en Staps (ou autre), la situation avec 3 choix est inutilement complexe (Gerville-Réache L. (2022)). L'approche 2x2 avec uniquement les choix « droite » et « gauche » pour le gardien et le tireur permet de pleinement introduire la théorie des jeux non coopératifs, simultanés, à sommes constantes (ou nulles). Les statistiques utilisées sont précisées dans les deux tableaux 1 et 2 suivants :

**Tableau 1 : Pourcentage de buts marqués**

| | Tireur | |
|---|---|---|
| Gardien | Gauche | Droite |
| Gauche | 60% | 80% |
| Droite | 90% | 50% |

**Tableau 2 : Nombre de situations du jeu**

| | Tireur | |
|---|---|---|
| Gardien | Gauche | Droite |
| Gauche | 117 | 95 |
| Droite | 85 | 75 |

On peut lire par exemple (tableau 1) que dans la situation où le gardien part à gauche et le tireur tire à droite, le but est marqué dans 80% des cas (il s'agit d'un pourcentage arrondi pour plus de simplicité). On comprend que, dans ce cas, les buts non marqués sont dus à des tirs non cadrés… Attention, lorsque que l'on parle de tir à gauche, il s'agit plus précisément d'un tir croisé. Même si le tireur est gaucher, un tir croisé sera aussi dit « tir à gauche ». On peut lire par exemple (tableau 2) que sur les 372 pénaltys recensés, 117 ont vu la situation gauche-gauche se réaliser.

# 2 Modélisation via la théorie des jeux

Ces deux tableaux de données vont nous suffire pour questionner la « rationalité » des joueurs. Mais pour cela, il va d'abord falloir définir le concept de matrice des gains et construire cette matrice.

**Tableau 3 : Matrice des gains du jeu du pénalty**

| Tireur / Gardien | Gauche | Droite |
|---|---|---|
| Gauche | 4 / 6 | 2 / 8 |
| Droite | 1 / 9 | 5 / 5 |

Par exemple, la situation gauche-gauche du tableau 3 rapporte 6 points au tireur et 4 points au gardien. Dans le tableau 1 des pourcentages de buts marqués, on pouvait lire que dans cette situation, 60% des pénaltys produisaient un but (et donc 40% un « non-but »). La matrice des gains est la traduction (en théorie des jeux) des statistiques de buts. Nous pouvons maintenant étudier le « jeu du pénalty ».

Il est alors temps des présenter aux étudiants les concepts de stratégie pure, de stratégie mixte, d'espérance de gain et d'optimisation de l'espérance de gain… L'objectif étant de sensibiliser l'étudiant à cette théorie et d'aller au bout du problème : (1) avec quelles probabilités le tireur et le gardien doivent-ils choisir, par exemple « gauche » ? (2) les statistiques sont-elles en adéquation avec cette théorie ?

Afin de justifier la pertinence de cette approche pour une situation réelle de pénalty, c'est principalement la simultanéité des choix qui doit être discutée. Est-il raisonnable de considérer

que les attitudes des gardiens et tireurs, les capacités à lire des trajectoires ou des intentions soient ignorées dans la modélisation ? C'est une question délicate mais il semble bien qu'à haut niveau, gardiens et tireurs fassent des choix aussi indépendants que possible de leurs perceptions ; « cela va trop vite… ». Aussi, assimiler un pénalty à un jeu à choix simultanés semble pertinent (c'est, en tous cas, le choix de Chiappori *et al.* (2002)). Avec cette considération, la probabilité de chaque situation (gauche-gauche, gauche-droite, droite-gauche et droite-droite) est simplement le produit des probabilités.

On note $p_T$ la probabilité que le tireur tire à gauche et $p_G$ la probabilité que le gardien parte à gauche (le « G » dans $p_G$ est pour « Gardien » et pas pour « gauche »). Voici les deux espérances de gain ($E_G$ pour le gardien et $E_T$ pour le tireur) :

$$E_G = 4p_G p_T + 2p_G(1-p_T) + 1(1-p_G)p_T + 5(1-p_G)(1-p_T)$$

$$E_T = 6p_G p_T + 8p_G(1-p_T) + 9(1-p_G)p_T + 5(1-p_G)(1-p_T)$$

Pour une lecture claire, par exemple la situation gauche-gauche arrive avec probabilité $p_G p_T$ et le gain pour le gardien sera de 4 et de 6 pour le tireur. Ceci est bien conforme à la matrice des gains du tableau 3. On peut alors représenter, entre autres, la surface de l'espérance de gain du gardien en fonction des probabilités $p_G$ et $p_T$ (voir figure 1). On retrouve par exemple que si le gardien part à droite avec probabilité 1 et le tireur tire à gauche aussi avec probabilité 1, le gain du gardien sera de 1 (situation D|G).

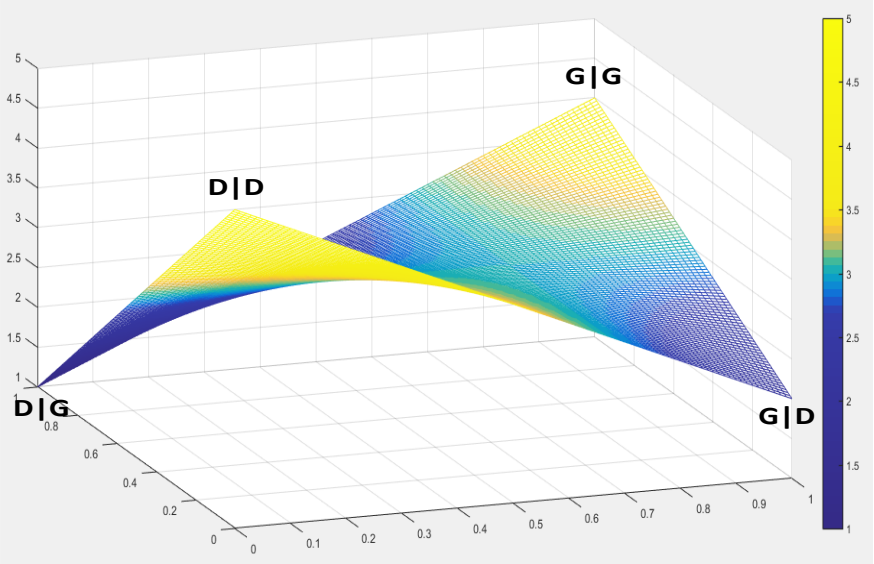



**Figure 1 : Espérance de gain du gardien en fonction de $p_G$ et $p_T$**

On pourrait aussi représenter la surface de l'espérance de gain du tireur mais comme le jeu est à somme constante (c'est-à-dire que dans chaque situation, la somme totale des gains est la même, ici 10) les surfaces sont complémentaires.

Dans l'article de Chiappori *et al.* (2002), il est question d'équilibre de Nash. Il s'agit pour chacun des joueurs de trouver *la meilleure réponse possible à la stratégie de l'autre*. Mais au jeu du pénalty, qui est un jeu à somme constante, ce point d'équilibre est identique à celui qui résulte simplement de « l'indépendance au choix de l'autre ». C'est-à-dire qu'il suffit pour chacun des deux joueurs de déterminer avec quelle probabilité il doit (par exemple) jouer « gauche » pour que son espérance de gain ne dépende pas de la stratégie de l'autre joueur. En plus d'être mathématiquement plus simple à déterminer, cette approche ne fait aucune hypothèse préalable sur la rationalité des joueurs et sa connaissance commune (il s'agit donc

de présenter davantage la théorie de Von-Neumann et Morgenstern que celle de Nash).

Plus intéressant encore, il est graphiquement assez simple de pressentir ces probabilités. Sur la figure 4, il semble que lorsque le gardien part à gauche avec probabilité 2/3, son espérance de gain ne varie plus en fonction de la stratégie du tireur. Réciproquement, lorsque le tireur tire à gauche avec probabilité 1/2, il semble alors que quoi que fasse le gardien, son espérance de gain ne varie pas non plus.

Vérifions avec l'équation de l'espérance de gain du gardien :

$$E_G = 4p_G p_T + 2p_G(1 - p_T) + 1(1 - p_G)p_T + 5(1 - p_G)(1 - p_T)$$

Avec $p_G = 2/3$, quelle que soit $p_T$, voici ce que l'on obtient :

$$E_G = 8/3p_T + 4/3(1 - p_T) + 1(1 - 2/3)p_T + 5(1 - 2/3)(1 - p_T) = 3$$

Et, avec $p_T = 1/2$, quelle que soit $p_G$, on trouve également que $E_G = 3$.

Nous y voilà, en partant à gauche avec probabilité 2/3, le gardien se garantit, quelle que soit la stratégie du tireur, une espérance de gain de 3 (et donc le tireur aura une espérance de gain de 7). Réciproquement, le tireur tirant à gauche avec probabilité 1/2, celui-ci se garantit une espérance de 7, quelle que soit la stratégie du gardien (voir figure 2).

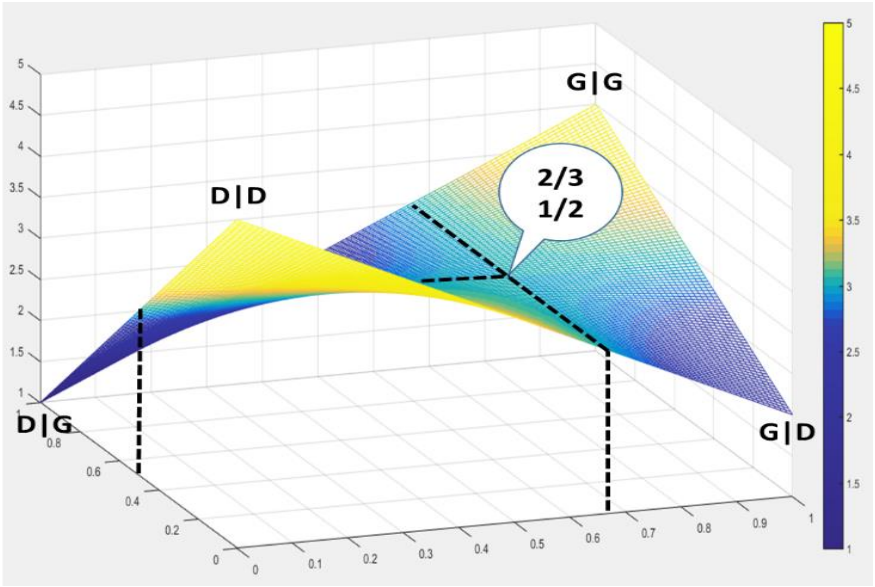



**Figure 2 : probabilités remarquables**

Il faut noter à ce stade que les joueurs cherchent bien à maximiser leurs espérances de gain mais que la seule option rationnelle réside dans la neutralisation de l'adversaire. Cette neutralisation garantit une espérance de gain constante minimax (ou maximin selon le point de vue).

# 3 Résultats

Nous avons la solution théorique du jeu. Maintenant, comparons-la aux statistiques de situations vues dans le tableau 2. La question est alors la suivante : idéalement, si gardien et tireur jouaient selon la théorie, à quoi ressemblerait le tableau de situations de 372 pénaltys ?

C'est particulièrement simple : par exemple, la situation gauche-gauche aurait une probabilité de 2/3 x ½ de se produire (soit 1/3). Idéalement, dans la case gauche-gauche, on aurait alors 124 fois la situation (1/3 x 372). Plus globalement, voici le tableau idéal théorique.

**Tableau 4 : Nombre de situations théoriques du jeu**

| Gardien | Tireur | |
|---|---|---|
| | Gauche | Droite |
| Gauche | 124 | 124 |
| Droite | 62 | 62 |

Maintenant il reste à comparer le tableau des observations avec le théorique pour voir si la différence (visible) est statistiquement significative. Pour cela, c'est un test du Khi-deux que l'on peut recommander. Ici la p-value vaut 0,000018. La différence entre les deux tableaux (l'observé et le théorique) est clairement significative. Cela signifie que les observations sont en désaccord avec la théorie. Il semble que les gardiens et/ou les tireurs ne suivent pas les recommandations de la théorie des jeux…

Dans l'article de Chiappori *et al.* (2002), la conclusion est différente et c'est l'occasion de revenir sur les éléments qui conduisent à un résultat significatif ou non. Dans l'étude originale, le choix est triple (avec la question du tir au centre) alors que dans l'adaptation pédagogique, la situation centrale a été retirée. De plus, les statistiques de la situation pédagogique ont été arrondies pour produire des calculs plus simples et une solution numérique tout aussi simple. Aussi, les p-values et les conclusions ne sont pas directement comparables et ne sont pas contradictoires.

# 4 Conclusion

L'analyse du jeu du penalty permet de mêler des statistiques sur les choix de gardiens et tireurs avec des statistiques de réussites et d'échecs. La matrice des gains qui peut être construite provient alors d'une réalité statistique qui encre le problème du penalty dans la réalité sportive. En sport, il existe bien d'autres situations où la théorie des jeux 2x2 ou 3x3 reste un outil de réflexion et d'analyse pertinent. La question du service au tennis est un exemple d'application classique mais certains chercheurs travaillent sur des jeux originaux pour mettre les sportifs ou des équipes dans des situations de dilemmes. Par exemple Dugas E, Collard L. (2009) propose une situation de jeu de ballon expérimentale avec la matrice des gains suivante :

**Tableau 5 : Matrice des gains (Dugas E, Collard L. (2009))**

| | | Équipe B | |
|---|---|---|---|
| | | ballon de HB | ballon de GR |
| Équipe A | ballon de HB | {2,3} | {3,1} |
| | ballon de GR | {4,2} | {1,4} |

# Bibliographie


Chiappori P.-A., Levt S., Groseclose T. (2002), *Testing Mixed-Strategy Equilibria When Players are Heterogene: The case of Penalty Kicks in Soccer*, American Economic Review (vol. 92), pp. 1138-1151.

Dugas E, Collard L. (2009), *Les sportifs et les interactions stratégiques sous l'angle de la théorie des jeux expérimentale*, Les Cahiers Internationaux de Psychologie Sociale (Numéro 81), pp. 7-24.

Gerville-Réache L. (2022), Statistique et traitement des données : du recueil à l'interprétation, *Edition Ellipses, collection Objectif STAPS*, 228p.

# La Science Statistique au service des Jeux Paralympiques 2024 : l'exemple du tir à l'arc

Christian Derquenne

*Chercheur indépendant - chris.emr@wanadoo.fr*

**Résumé.** Cet article a pour objectif de montrer en quoi la Science Statistique peut répondre aux enjeux issus du domaine du Sport. Nous illustrons cette dimension dans le domaine para-sportif et plus particulièrement axée vers les Jeux Paralympiques de 2024 à Paris. Les travaux présentés entrent dans le cadre du projet Paraperf de l'INSEP et du Mécénat de compétences EDF appuyé par EDF R&D. La discipline sportive présentée dans cet article est le tir à l'arc. Nous analysons les liens entre les phases qualitatives et finales, nous proposons des indicateurs de performance, et nous les utilisons pour construire des typologies d'archers et hiérarchiser leurs contributions au gain en finale.

**Mots-clés.** Science du Sport, Science Statistique, Jeux Paralympiques, tir à l'arc, indicateurs statistiques de performance individuelle

**Abstract.** This article aims to show how Statistical Science can respond to issues arising from the field of Sport. We illustrate this dimension in the para-sports field and more particularly focused on the Paralympic Games which will take place in Paris in 2024. The work presented falls within the framework of the Paraperf project of INSEP and the EDF Skills Sponsorship supported by EDF R&D. TThe sporting discipline presented in this article is archery. We analyze the links between the qualitative and final phases, we propose performance indicators, and we use them to construct typologies of archers and prioritize their contributions to winning in the final.

**Keywords.** Science of Sport, Statistical Science, Paralympic Games, archery, statistical indicators of individual performance

## 1 Contexte - objectif

2024 est une année phare pour le multi-sport : les Jeux Olympiques et Paralympiques (JOP) d'été en sont les dignes représentants. Ces deux événements majeurs organisés tous les quatre ans représentent sans doute le plus grand rassemblement de sportives et de sportifs de haut niveau réparti(es) en valides ou en situation de handicap au sein d'une multitude de disciplines les plus diverses. C'est une des raisons de leur succès, mais l'aventure humaine est la véritable raison d'être. En effet, les JOP permettent à la plupart des nations du Monde de se rencontrer avec un esprit de paix et de bonne intelligence. Bien évidemment la performance sportive et son amélioration représentent la clé de voûte des JOP comme toutes les compétitions sportives. Outre, les aspects physique, technique, stratégique, tactique, psychologique, matériel qui composent la performance sportive, la recherche de celle-ci s'appuie de plus en plus sur la Science (physique, chimie, biologie, sociologie, mathématiques, ...). Les britanniques ont notamment fait appel à cette dernière pour augmenter de façon significative le nombre de médailles lors des JOP

de Londres en 2012. Cette pluridisciplinarité de la Science est par conséquent de plus en plus utilisée par des fédérations et des clubs sportifs par le biais d'analyses poussées de la captation vidéo, d'amélioration du matériel, d'entrainements adaptés tenant compte de la charge physique et mentale pour éviter les blessures, ...

Cet article est dédié aux Jeux Paralympiques. En effet la situation de handicap fait partie de notre vie de tous les jours, qu'il soit physique, mental, de naissance ou accidentelle. L'intégration pleine et entière du handicap dans le monde sportif représente un lien humain et social fort en termes de partage et de sens de la vie. L'accès de personnes en situation de handicap à un nombre croissant de disciplines sportives en est la preuve : organisation d'événements sportifs allant des amateurs aux professionnels en rassemblant de plus en plus de types de handicap. Par conséquent, les JP 2024 constituent un enjeu fort en termes de reconnaissance sportive du handicap. Les parathlètes françaises et français sont les premiers acteurs avec leurs coaches pour réussir. L'utilisation de la Science est là pour les aider dans leurs stratégies et leurs décisions. L'INSEP a lancé en 2019 un vaste programme de fundraising comportant un volet mécénat et un volet sponsoring. Ce programme est conçu autour de valeurs humaines, orientées particulièrement vers l'accompagnement des athlètes. C'est dans ce cadre qu'EDF et plus particulièrement sa R&D a offert ses compétences scientifiques dans le cadre du Mécénat de compétences en contribuant au projet Paraperf (Optimisation de la performance Paralympique : de l'identification à l'obtention de la médaille). Celui-ci capitalise sur une approche interdisciplinaire ; il doit optimiser le parcours de chaque athlète français visant un podium aux Jeux Paralympiques de Paris 2024. Ce projet possède trois axes majeurs : les trajectoires de performance des parathlètes français(e)s et de leurs adversaires, les équipements, et les aspects socio-facteurs environnementaux. La présente communication traite du premier axe sous l'angle de la performance sportive dans les compétitions de haut niveau à l'aide d'indicateurs statistiques individuels et collectifs. L'objectif de ce lot est de comprendre et de situer les niveaux de performance des athlètes dans un contexte concurrentiel inter-épreuves. Cela peut aider à déterminer où les écarts relatifs sont les plus importants, ce qui révèle de plus grandes opportunités. En 2023, deux études de ce projet sur la densité de courses para-cyclistes sur route et sur la construction d'indicateurs de performance pour le développé-couché en haltérophilie ont fait l'objet d'un article [Derquenne, 2023]. Dans le présent papier, nous traitons de différentes problématiques sur le tir à l'arc dans la section 2. Nous concluons sur de futures études dans le cadre du projet Paraperf.

## 2 Etude sur le tir à l'arc pour les Jeux Paralympiques 2024

### 2.1 Problématiques, données et objectifs

Les compétitions en tir à l'arc se déroule en deux phases : une phase de qualification et une phase de confrontation (phase finale). Dans la phase de qualification, chaque archer tire 72 flèches ou 144 flèches selon la compétition, la somme des points constitue le classement et permet de construire l'arbre des confrontations. Un athlète bien classé va affronter un athlète moins bien classé (parfois, les premiers passent directement le 1er tour en fonction du nombre de participants). Le nombre de rounds varie en fonction du nombre de participants qui peut

être différent en fonction des catégories, du genre et de la compétition. Dans la phase de confrontation, pour chaque set, chaque archère ou archer tire 3 flèches, la somme des 3 tirs est effectuée, alors, le gagnant obtient 2 points, le perdant, 0 et 1 partout en cas d'égalité). Il y a 5 sets maximum (Le premier à 6 points remporte le match). En cas d'égalité au 5ème set : mort subite, ils tirent 1 flèche, celui qui met le plus de points remporte le match.

Les informations sont issues de données scrapées sur le site : https://worldarchery.sport/fr, à l'aide de l'API : https://api.worldarchery.org/v3/API/. Ces données contiennent toutes les compétitions depuis 2009 : 46 compétitions avec confrontations ; 4 593 confrontations individuelles ; 44 compétitions avec qualifications ; 243 épreuves.

Le principal objectif fixé par la DTN de la fédération du tir à l'arc est d'étudier le lien entre les résultats de la phase de qualification et de la phase de confrontation. Cependant, il s'agit d'un sujet large, il n'y avait pas une demande précise. Par conséquent, toute la liberté est donnée pour compléter cette demande en termes de méthodologies et d'approches statistiques.

Nous avons mis oeuvre plusieurs axes de travail : (i) Lien entre les rangs obtenus en qualification et en confrontation (ii) Construction d'indices de performance individuelle (iii) Typologie des par-athlètes à l'aide des indicateurs de performance (iv) Analyse multidimensionnelle des indicateurs de performance (v) Approches pour prédire l'accès au Podium et en Top 8 (vi) Etude des confrontations à l'aide des indices de performance (vii) Comparaison du classement réel de la compétition et de l'indicateur global de performance individuelle (viii) Classement annuel global des athlètes, des pays et des compétitions en fonction d'indices de performance (ix) Prévision de performances futures à l'aide performances passées.

Chacune des deux phases a sa propre table de données. ***La phase de qualification*** possède les variables suivantes : rang du par-athlète en qualification, nombre de points obtenus (72 ou 144 flèches), nombre de 10 et de X (perfect, plein centre de la cible), nombre de X, identifiant du par-athlète, nom du par-athlète, prénom du par-athlète, nationalité du par-athlète, catégorie du handicap (CMO/CWO : arc à poulie ; RMO/RWO : arc classique ; MW1/WM1 : fauteuil - handicap plus important) et le numéro de la compétition. ***La phase de confrontation*** contient le numéro de la compétition, le nom de la compétition, le lieu de la compétition, le pays de la compétition, la date de début de la compétition, la date de fin de la compétition, le numéro de la phase de confrontation, le libellé de la phase de confrontation, l'identifiant du par-athlète 1, son nom, son prénom, sa nationalité, sa catégorie de handicap, son rang final en confrontation, son rang en qualification, ses résultats des flèches sur les sets, ses résultats des flèches en mort subite, son nombre de points obtenus, son état gagnant/perdant, idem pour le par-athlète 2, son adversaire. Enfin, cette table de données a été transformée de façon à obtenir un résultat individuel pour chaque par-athlète à la place de deux résultats associés à la confrontation des deux par-athlètes.

### 2.2 Lien entre les rangs obtenus en qualification et en confrontation

L'objectif de cette étude est d'analyser le lien entre les rangs individuels en qualification et en confrontation en répondant à la question : "est-ce que plus un par-athlète est bien classé en phase de qualification, plus il a de chance d'obtenir un bon classement en phase finale ?".

Pour répondre à cette question, nous avons calculé des tables de correspondances ou de passage du rang en qualification au rang final. Pour un même type de handicap, nous avons rassemblé

tous les résultats de chacun des archers aux différentes compétitions auxquelles il a participées. Pour cela, nous avons construit un tri croisé entre les deux rangs.

Par exemple, pour les CWO, nous obtenons la table de correspondances (table 1) dans laquelle les lignes correspondent aux rangs de la finale ; les colonnes aux rangs de qualification. Par exemple, il y a eu 16 archères qui ont été classées 1ières sachant qu'elles ont également été classées 1ières en qualification (cf. 1ière ligne de la 1ière colonne), puis 7 archères ont obtenu la 1ière place en finale, sachant qu'elles étaient classées 2ièmes en phase de qualification, (cf. 1ière ligne de la 2ième colonne), etc. Ces résultats permettent de construire un diagramme à bâtons (figure 1) affichant les proportions d'être vainqueur en finale sachant le rang en qualification. Par exemple, elle est égale à 0,39 si l'archère a terminé au premier rang des qualifications, et tombe à 0,20 pour le rang 2, à 0,17 pour le rang 3, etc. Les lignes rouges horizontales pour chaque rang en qualification désignent la proportion observée de celui-ci. Par exemple, il y a une proportion de 0,06 d'archères qui ont obtenu le premier rang en qualification. Nous pouvons tout d'abord comparer empiriquement la proportion de rang 1 en qualification qui ont été vainqueurs et la proportion globale de rang 1 en qualification : 0,39 *vs* 0,06 ; elle est plus de six fois plus élevée.

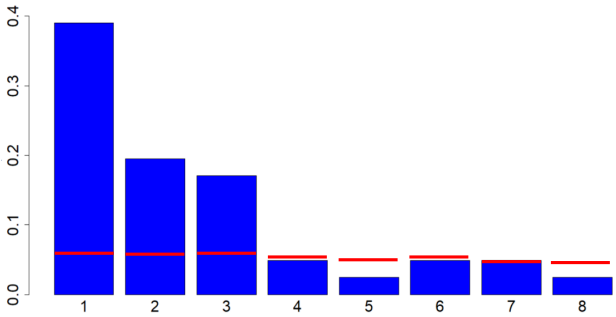



*Figure 1 : Proportions de gagner la finale en fonction du rang en qualification*

| F/Q | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 17 | 33 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 16 | 7 | 6 | 5 | 3 | 0 | 0 | 0 | 4 | 0 | 0 | 41 |
| 2 | 8 | 9 | 7 | 8 | 3 | 3 | 1 | 0 | 2 | 0 | 0 | 41 |
| 3 | 7 | 9 | 7 | 4 | 5 | 3 | 2 | 1 | 3 | 0 | 0 | 41 |
| 4 | 2 | 7 | 4 | 6 | 5 | 4 | 6 | 2 | 4 | 0 | 0 | 40 |
| 5 | 1 | 4 | 8 | 4 | 7 | 4 | 1 | 2 | 6 | 2 | 0 | 39 |
| 6 | 2 | 2 | 1 | 5 | 2 | 13 | 2 | 4 | 7 | 0 | 0 | 38 |
| 7 | 2 | 0 | 4 | 2 | 1 | 3 | 4 | 5 | 12 | 1 | 0 | 34 |
| 8 | 1 | 0 | 2 | 1 | 7 | 2 | 4 | 4 | 13 | 0 | 0 | 34 |
| 9 | 2 | 3 | 2 | 4 | 10 | 9 | 11 | 10 | 138 | 31 | 0 | 220 |
| 17 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 2 | 30 | 117 | 9 | 162 |
| 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 37 | 46 |
| Total | 41 | 41 | 41 | 40 | 43 | 42 | 33 | 30 | 219 | 160 | 46 | 736 |

*Table 1 : Tri croisé des rangs en qualification vs en confrontation*

La validation de ces comparaisons de proportions peut être réalisée à l'aide d'un test statistique dans lequel le jeu d'hypothèses est le suivant :

$H_0$ : Les $n^{(c_r)}$ individus sont tirés au hasard (sans remise) parmi les $N$ individus de la population

$H_1$ : La proportion de la modalité $q_r$ est "anormalement" élevée parmi les $n^{(c_r)}$ individus

où $N$ et $n^{(c_r)}$ sont respectivement le nombre total d'individus (ici 736) et le nombre d'archères ayant obtenu le rang $r$ en phase de confrontation (pour $r = 1$, nous avons 41).

Lorsque les marges d'un tableau de contingence sont fixées (les nombres de modalités pour les variables catégorielles sont connues), alors chaque effectif croisé $n_{rs}$ de cette table suit une loi Hypergéométrique $\mathcal{H}(N, n^{(c_r)}, n^{(q_s)})$ sous $H_0$, où $n^{(q_s)}$ est le nombre d'archères ayant atteint le rang $s$ en qualification (pour $s = 1$, nous obtenons 41). La $p$-valeur du test est donnée par $\Pr[\mathcal{H}(N, n^{(c_r)}, n^{(q_s)}) > n_{rs}^{obs}/H_0]$, où $n_{rs}^{obs}$ est le nombre observé d'individus dans la case $(r, s)$ du tri croisé, par exemple 16 dans la table 1 pour la case $s = 1$ et $r = 1$.

Dans [Morineau, 1982], l'auteur propose de construire une valeur-test égale au fractile de la loi normale centrée réduite $\mathcal{N}(0; 1)$ au point $1 - p$-valeur. Plus la $p$-valeur est faible, plus la

valeur-test est hautement positive et plus une modalité est typique dans sa classe. A l'opposé, une $p$-valeur élevée indiquera une modalité rare. En pratique, la valeur-test doit dépasser 2 à 3 en valeur absolue pour considérer que la modalité est fréquente ou rare de manière significative.

La table 2 permet d'identifier les rangs de qualification typiques de la place de vainqueur en finale. Les résultats issus des tests confirment les résultats empiriques. En effet, le rang 1 en qualification est très fréquent : 39,02% *vs* 5,57%, la $p$-valeur est très faible et fournie une valeur-test égale à 6,74 donc significatif. Comme les rangs 2 et 3 en qualification qui obtiennent respectivement des valeurs-test de 3,21 et de 2,72. Les autres rangs ne sont pas significatifs ni de façon fréquente, ni de façon rare.

| Rang | Mod/Cla | Global | p.value | v.test |
|---|---|---|---|---|
| Qr=1 | 39.02 | 5.57 | 1.60e-11 | 6.74 |
| Qr=2 | 19.51 | 5.57 | 1.33e-03 | 3.21 |
| Qr=3 | 17.07 | 5.57 | 6.50e-03 | 2.72 |
| Qr=7 | 4.88 | 4.62 | 8.70e-01 | 0.16 |
| Qr=6 | 4.88 | 5.16 | 9.96e-01 | -0.01 |
| Qr=4 | 4.88 | 5.43 | 9.43e-01 | -0.07 |
| Qr=8 | 2.44 | 4.62 | 5.58e-01 | -0.59 |
| Qr=5 | 2.44 | 5.30 | 4.46e-01 | -0.76 |

*Table 2 : Tests de comparaison de proportions pour les vainqueurs*

Les rangs 2, 3, puis 4 et 1 en qualification pour les finalistes sont significativement fréquents, alors que les rangs 5, 2, 3 et 1 le sont pour la $3^{ieme}$ place du podium. Cette corrélation "positive" de bons rangs en confrontation avec les rangs en qualification se poursuit pour les cinq suivants. Mais ce résultat fort pour la catégorie CWO est plus modulé pour les archers CMO.

### 2.3 Construction d'indices de performance individuelle

La mesure de performance individuelle fait partie des éléments important pour les coaches dans les fédérations car ils leurs permettent de disposer d'information pour les aider dans leurs décisions. Le principe pour construire des indices de performance individuelle est d'identifier des informations discriminantes, compréhensibles et surtout utilisables par les coaches, dans les données permettant de mesurer certaines capacités sportives du par-athlète. Deux types d'indicateurs ont été construits : statiques et dynamiques. Le premier type est associé aux résultats pour chacune des deux phases (qualification et confrontation), alors que le second mesure l'évolution des résultats entre les deux phases.

#### *2.3.1 Indicateurs statisques*

Sept indices statiques ont été construits.

***Le premier indice*** met en avant le nombre de "X" (perfect) obtenus par chaque archer lors de la phase de qualification (1).

$$I_{qX}(i) = \frac{\sum_{m=1}^{M_q} 1_{[f_m=X]}}{M_q} \quad (1)$$

où $M_q$ est le nombre de flèches tirées, $f_m$ est le score obtenu pour la $m^{ieme}$ flèche. $1_{[f_m=X]} = 1$ si la $m^{ieme}$ flèche est au centre de la cible (X), sinon $1_{[f_m=X]} = 0$. Par conséquent cet indicateur représente la proportion de "perfects". Plus la valeur est élevée, plus l'archer est performant pour atteindre le centre de la cible.

***Le deuxième indice*** repose sur la différence de rangs entre deux compétiteurs. On pose tout d'abord cette différence : $b_{(i/j)} = r_{(q(i))} - r_{(q(j))}$ où $r_{(q(i))}$ et $r_{(q(j))}$ sont respectivement les rangs de qualification des archers $i$ et $j$. L'objectif de cet indicateur est de bonifier le par-athlète $i$ si son rang est plus élevé que celui de $j$, s'il a gagné la confrontation (2.1), alors que s'il perd et que son rang est plus faible que son adversaire, alors il est pénalisé (2.4). Par contre, cet indicateur est égal à 0, s'il gagne alors que son rang était meilleur que son adversaire (2.3), ou s'il perd alors que son rang était moins bon que son adversaire (2.2).

$$\text{Si } b_{(i/j)} > 0 \text{ et } W_i = 1 \text{ alors } I_{dr}(i) = b_{(i/j)} \tag{2.1}$$

$$\text{Si } b_{(i/j)} > 0 \text{ et } W_i = 0 \text{ alors } I_{dr}(i) = 0 \tag{2.2}$$

$$\text{Si } b_{(i/j)} < 0 \text{ et } W_i = 1 \text{ alors } I_{dr}(i) = 0 \tag{2.3}$$

$$\text{Si } b_{(i/j)} < 0 \text{ et } W_i = 0 \text{ alors } I_{dr}(i) = -b_{(i/j)} \tag{2.4}$$

où $W_i = 1$ représente le gain du match, alors que la perte est donnée par $W_i = 0$.

Cet indicateur est symétrique par rapport à son adversaire $j$. En effet, si $I_{dr}(i) = b_{(i/j)}$ alors $I_{dr}(j) = -b_{(i/j)} = b_{(j/i)}$ ; si $I_{dr}(i) = -b_{(i/j)}$ alors $I_{dr}(j) = b_{(i/j)} = -b_{(j/i)}$ et si $I_{dr}(i) = 0$ alors $I_{dr}(j) = 0$. Par conséquent, une valeur positive indiquera un gain notable de performance par rapport à son adversaire, alors que valeur négative engendrera une sorte de manque à gagner.

***Le troisième indice*** mesure le score moyen des scores obtenus lors de la phase de confrontation par un archer $i$ sur la même compétition, tel que :

$$I_{(sc_cf)}(i) = \frac{\sum_{m=1}^{M_c} f_m}{M_c} \tag{3}$$

avec $M_c$ le nombre de flèches tirées en phase de confrontation (12 par match). Plus la valeur observée de (3) sera grande, plus l'archer sera performant.

***Le quatrième indice*** utilise le rang $r_{(f(i))}$ obtenu par l'archer $i$ lors de la phase finale :

$$I_{rang}(i) = r_{f(i)} \tag{4}$$

Contrairement aux indices introduits précédemment pour lesquels une grande valeur signifie une performance élevée, plus une valeur de (4) est faible, plus la performance de l'archer est forte.

***Le cinquième indice*** mesure le nombre de "X" (perfect) obtenus pour chaque archer lors de la phase de confrontation (5).

$$I_{fX}(i) = \frac{\sum_{m=1}^{M_f} 1_{[f_m=X]}}{M_f} \tag{5}$$

où $M_f$ est le nombre de flèches tirées sur l'ensemble des confrontations. Comme pour (1), cet indicateur représente la proportion de "perfects". Plus la valeur est élevée, plus l'archer est performant pour atteindre le centre de la cible.

***Le sixième indice*** correspond à la dispersion des scores obtenus sur les jeux de 12 ou 15 flèches lors de la phase de confrontation. Il permet de mesurer la régularité (l'homogénéité ou la reproductibilité) des tirs de l'archer $i$.

$$I_{disp}(i) = \sqrt{\frac{\sum_{m=1}^{M} (f_m - \bar{f})^2}{M-1}} \tag{6}$$

Comme l'indice sur le rang final (4), plus une valeur de (6) est petite, plus le par-athlète est régulier, ce qui peut être signe d'un meilleur contrôle de soi.

Le septième indice correspond au coefficient de variation des scores obtenus sur les jeux de 12 ou 15 flèches lors de la phase de confrontation. Il permet de mesurer la dispersion par rapport à la moyenne des points des tirs de l'archer $i$.

$$I_{cv}(i) = \frac{I_{disp}(i)}{\bar{f}} \tag{7}$$

Plus une valeur de (7) est petite, plus le par-athlète est régulier, ce qui peut être signe d'un meilleur contrôle de soi comme dans (6) mais aussi accompagné d'un score élevé.

*2.3.2 Indicateurs dynamiques*

Les trois indices mesurent l'évolution des résultats de chaque archer entre les 2 phases.

***Le premier indice*** repose sur la différence de rangs entre les deux phases pour chaque archer (8). Si la valeur est positive alors l'archer $i$ a amélioré son classement, si elle est négative, il a déprécié son rang, si elle est nulle, il a fait ce qui était attendu. On peut interpréter cet indice comme une évolution (positive, négative ou nulle) de son classement par rapport à une espérance qui correspond au rang de l'archer en phase de qualification.

$$I_{(ev-rg)}(i) = r_{q(i)} - r_{f(i)} \tag{8}$$

où $r_{q(i)}$ et $r_{f(i)}$ sont respectivement les rangs obtenus en qualification et en confrontation.

***Le deuxième indice*** mesure l'évolution du score moyen (8) entre les deux phases :

$$I_{ev-sc}(i) = I_{sc-cf}(i) - I_{sc-ql}(i) \tag{9}$$

où $I_{(sc-ql)}(i) = \frac{\sum_{m=1}^{M_q} f_m}{M_q}$, avec $M_q$ le nombre de flèches tirées en phase de qualification. Plus une valeur de (9) est grande, plus l'archer a amélioré son score moyen entre les deux phases.

***Le troisième indice*** calcule la différence entre les moyennes des perfects (X) obtenus lors des deux phases. Plus une valeur de (10) est grande, plus l'archer a augmenté sa précision.

$$I_{ev_X}(i) = I_{fX}(i) - I_{qX}(i) \tag{10}$$

Les 10 indices introduits précédemment prennent des valeurs sur des échelles différentes. En effet, les rangs ou leur différence ne sont pas comparables aux scores moyens ou à leur différence, ni à la dispersion des flèches, etc. Par conséquent, une transformation de ces indices est nécessaire pour les rendre comparables et par la suite construire des indicateurs globaux. Pour cela, tous les indices sont transformés afin qu'ils prennent leurs valeurs entre 0 et 1. Plus le résultat de l'indicateur sera proche de 1, plus la performance de l'archer sera élevée. Pour les indicateurs de dispersion et de coefficient de variation intrinsèques, une valeur proche de 1, indiquera la régularité des performances de l'archer. La transformation appliquée est la suivante :

$$I^*(i) = \frac{I(i) - \min\limits_j I(j)}{\max\limits_j I(j) - \min\limits_j I(j)} \tag{11}$$

Ces 10 indices permettent de réaliser de nombreuses études comme indiquées dans la partie 2.1. Nous avons choisi de présenter deux études une typologie des par-athlètes et l'analyse de l'importance de contribution des indices de performance dans le gain du match.

### 2.4 Typologies des par-athlètes à l'aide des indices de performance

Afin d'enrichir ces analyses, il peut être fructueux de rechercher des structures dans les données en construisant une typologie des par-athlètes à l'aide d'informations les renseignant. Dans le cas présent, nous utiliserons l'ensemble des indicateurs de performance individuelle, mais il serait également possible de séparer les indicateurs statiques et dynamiques. La méthode de classification est le critère de Ward adapté à des données numériques.

L'arbre de classification (dendrogramme) montre plutôt cinq classes (figure 2).

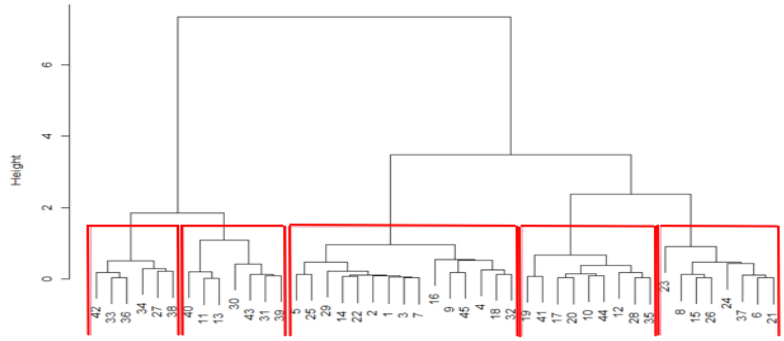



*Figure 2 : Arbre de classification des archers*

L'interprétation des classes est réalisée à l'aide du concept de la valeur-test permettant de classer l'importance des variables initiales ou supplémentaires et de déterminer le sens de celles-ci pour chaque groupe. Tout d'abord, les résultats suivants (table 3) montrent que les indicateurs les plus structurants pour construire la typologie sont les indices globaux statique et dynamique, le rang final avec des rapports de corrélation de plus 0,75 (plus une valeur est proche de 1, plus la variable est structurante), puis le score moyen final (0,61). Enfin, les trois indicateurs dynamiques possèdent également un poids important dans construction de la classification (0,57).

$$\rho^2(X_j, T) = \frac{\sum_{m=1}^{M_f} n_m(\bar{X}_{jm} - \bar{X}_j)^2}{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2} \tag{12}$$

où $M$, $n_m$, $n_m$ et $\bar{X}_j$ sont respectivement le nombre de classes de la typologie T, le nombre d'individus de la classe $m$, $\bar{X}_{jm}$ la moyenne de $X_j$ associée et la moyenne de $X_j$ dans l'échantillon.

| ***Indice*** | $I^*_{sta/dyn}$ | $I^*_{rang}$ | $I^*_{sc_cf}$ | $I^*_{ev_X}$ | $I^*_{ev_rg}$ | $I^*_{ev_rg}$ | $I^*_{fX}$ | $I^*_{disp}$ | $I^*_{qX}$ | $I^*_{dr}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\rho^2(X_j, T)$ | 0,765 | 0,756 | 0,609 | 0,577 | 0,576 | 0,571 | 0,561 | 0,494 | 0,485 | 0,425 |

*Table 3 : Importance des indices de performance*

Le sens et la force des variables sont fournis par le concept de valeur-test introduit dans le paragraphe 2.2, mais cette fois-ci dans le cas de variables numériques, tel que :

$$vt(x_j) = \frac{\bar{x}_{jm} - \bar{X}_j}{s_{jm}} \tag{13}$$

où $s^2_{jm} = (n - n_m)/(n-1)s^2_j$ et $s^2_j$ est la variance de la variable $X_j$

Elle s'interprète comme un nombre d'écarts-types autour de la moyenne. Plus la valeur-test est positive, (resp. négative), plus la moyenne $\bar{X}_{jm}$ de la variable $X_j$ dans la classe $m$ est éloignée supérieurement (resp. inférieurement) de la moyenne $\bar{X}_j$ de la variable $X_j$ sur l'ensemble de l'échantillon. Dans notre cas, nous avons fixé des valeurs-tests limites à -1,96 ou 1,96.

Pour ***la classe 1*** (figure 4), le nombre moyen de perfects en phase de qualification est l'indice le plus discriminant et fortement supérieur à la moyenne générale pour l'ensemble des archers sur cette compétition. La valeur-test est égale à 4,19 (4,2 écarts-types), la moyenne dans la classe vaut 0,551, alors que la moyenne globale est égale 0,384. Par contre, la moyenne de l'indicateur de l'évolution du nombre moyen de X est plus faible que la moyenne générale associée (0,316 *vs* 0,515), la valeur-test est égale à -3,72. Les indices sur le rang, du score moyen et de la dispersion en finale possèdent des moyennes plus élevées que les moyennes globales associées. Ce groupe contient des par-athlètes performants pour les indices associés aux valeurs-tests positives.

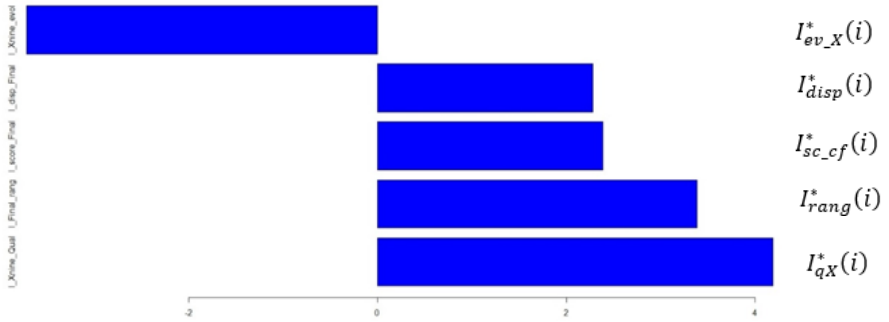



*Figure 4 : Représentation des valeurs-tests pour la classe 1*

***La classe 2*** possède des archers très performants sur de nombreux indicateurs, en particulier ceux d'évolution. La différence des rangs entre les deux archers en confrontation, les évolutions du score moyen et des rangs des archers entre les deux phases apparaissent de façon positive dans ***la classe 3***. Par contre, le nombre moyen de plein centre dans la cible, lors de la phase qualificative, apparait comme plus faible que la moyenne générale. Cette classe contient également des archers performants. ***La classe 4*** possède des par-athlètes moins performants que la moyenne générale pour de nombreux indicateurs. Seul l'indice sur l'évolution des perfects entre les deux phases apparait de façon positive. Enfin, tous les indices de la ***classe 5*** possèdent des moyennes plus faibles que les moyennes générales associées. Les par-athlètes sont peu performants.

### 2.5 Importance des indices de performance individuelle sur la victoire

L'objectif est d'évaluer globalement par compétition, la force d'explication des indicateurs sur le gain du match. Pour cela, nous utilisons la valeur-test. Dans le cas présent les deux catégories à "expliquer" sont la victoire et la défaite. Dans ce cas, les valeurs-tests fournissent l'ordre d'importance des indices de performance individuelle. Elles peuvent être représentées sous forme d'un diagramme à bâtons plus lisible qu'un tableau de résultats (figure 5). La longueur de chaque barre correspond à la valeur-test associée à chaque indice. Elles sont toutes du côté positif. Les lignes verticales du milieu représentent la limite de "significativité" des indices. Toutes les

valeurs-tests qui dépassent la troisième ligne verticale correspondent aux indices significatifs. ***I-Final-rang*** est l'indice qui agit le plus sur la victoire de façon positive. La moyenne de celui-ci est de 0,844 pour la catégorie des gagnants alors qu'il vaut 0,723 sur l'ensemble de l'échantillon.

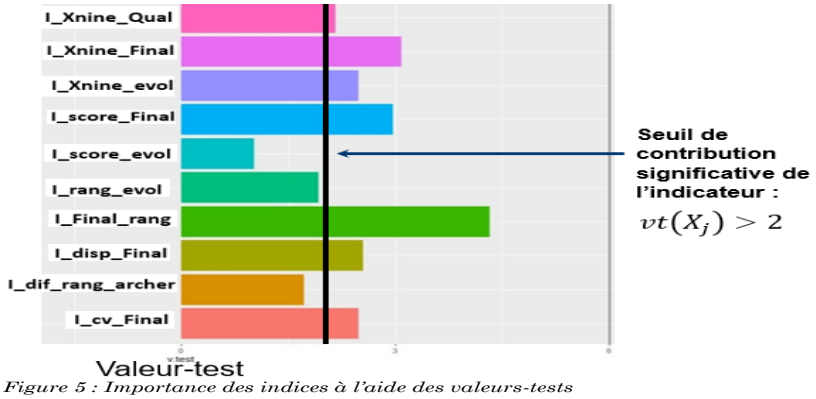



*Figure 5 : Importance des indices à l'aide des valeurs-tests*

## 3 Synthèse des résultats et voies futures

Les résultats présentés dans cet article ont montré que pour l'ensemble des compétitions les corrélations entre les rangs des deux phases étaient élevées, quel que soit le type d'handicap. Nous avons construit des indices de performance individuelle. Grâce à ceux-ci, nous avons recherché des structures en mettant oeuvre une analyse typologique pour obtenir des groupes d'archers et identifier leurs caractéristiques, ainsi que hiérarchiser les indicateurs de performance. D'autres analyses, non présentées ici, ont permis de détecter les caractéristiques potentielles pour atteindre le Top 8. Mais l'objectif majeur de cet article, outre la mise en oeuvre de méthodologies statistiques, était de montrer l'importance de la médiation scientifique, en d'autres termes la traduction des résultats obtenus en "langage terrain", ainsi que la mise à disposition d'outils d'aide à la décision à destination des fédérations, des coaches, des athlètes permettant d'adopter des stratégies adaptées aux parathlètes de façon individuelle et à l'égard de la concurrence. Cette chaine vertueuse offre un gain de temps, mais les décisions restent à la main des professionnels du sport. Enfin, nous développerons d'autres indices de performance individuelle et nous analyserons dans le temps l'évolution de différents par-athlètes.

## Bibliographie

Derquenne Ch., (2023), Les Jeux Paralympiques 2024 : la Science Statistique en action, *54ièmes Journées de Statistique*, Bruxelles, Belgique.
Morineau A., (1987), Inferential techniques folowing a Multivariate Descriptive Statistical Analysis, *Actes du congrès d'Analyse des Données en Barcelone.*
Projet Paraperf, (2020), INSEP.

# ESTIMATION DE TRAJECTOIRE ET ESTIMATION DE POTENTIEL DANS LES SPORTS PARALYMPIQUES

Imad Hamri[1,4], Julien Schipman[1] , Mélanie Baconnais[1], Bryan Le toquin[1,2,3], Nicolas Fortsman[1], Christian Derquenne[4]

[1]*Institut de Recherche bio-Médicale et d'Épidémiologie du Sport (IRMES), UPR 7329, Institut National du Sport, de l'Expertise et de la Performance (INSEP)*

[2]*Université de Paris, Paris, France*

[3]*Federation Française Handisport, France*

[4]*Société française de statistique, France*

**Résumé.**

PARAPERF est un programme de recherche prioritaire, financé par le dispositif France 2030, en partenariat avec la Fédération Française d'Handisport (FFH), la Fédération Française de Badminton (FFBad) et la Fédération française de Tir (FFTir). L'objectif principal de ce programme est de soutenir les athlètes et leur staff dans leur quête de performance et de médailles aux Jeux Paralympiques de Paris 2024. Le projet se structure en trois lots, dont le premier est dédié à l'estimation de potentiel et aux trajectoires de performance. Ce lot de travail utilise principalement les données de performance réalisées en compétition afin d'établir des études statistiques et des outils d'aide à la décision à destination des différents staff des équipes de France.

Des études sur la modélisation de la relation "âge-performance" sont réalisées pour de nombreuses disciplines telles que le para athlétisme, le para-cyclisme ou la para-natation, permettant ainsi la création de "couloirs de performance". Ces derniers servent à analyser les trajectoires de performance des athlètes et à estimer leur potentiel face à la concurrence internationale. Ces études permettent d'apprécier la dynamique de chaque athlète afin de voir leur évolution de manière mesurée et objective.

Pour les sports de confrontation, tels que l'escrime fauteuil, le para tir sportif ou le para tir à l'arc, des modèles de rating sont développés pour évaluer objectivement le niveau des athlètes. Des indicateurs statistiques propres à chaque sport sont calculés et fournis aux équipes pour optimiser la préparation de leurs entraînements et leur stratégie en compétition.

Dans des sports d'équipe comme le basket fauteuil, des indicateurs individuels et collectifs sont calculés afin d'objectiver les performances et d'analyser les points faibles et forts de chaque équipe. De nombreux éléments sont étudiés tels que la stratégie liée au niveau d'handicap dans la composition d'équipe ou l'efficacité des tirs en fonction des zones du terrain.

Le projet couvre 15 différents sports paralympiques ce qui permet une grande richesse d'analyse avec de nombreuses problématiques et méthodologies mises en place afin d'accompagner les différents acteurs qui préparent les Jeux Paralympiques de Paris 2024.

**Mots-clés :** Jeux Paralympiques, Haute performance, Analyse de données

**Abstract.**

PARAPERF is a priority research program, funded by the France 2030 initiative, in partnership with the French Handisport Federation (FFH), the French Badminton Federation (FFBad), and the French Shooting Federation (FFTir). The main goal of this program is to support athletes and their staff in their pursuit of performance and medals at the Paris 2024 Paralympic Games. The project is structured in three parts, the first of which is dedicated to potential estimation and performance trajectories. This work package primarily uses competition performance data to establish statistical studies and decision-making tools for the different French team staffs.
Studies on modeling the "age-performance" relationship are conducted for many disciplines such as para athletics, para cycling, or para swimming, thus creating "performance corridors". These are used to analyze athletes' performance trajectories and estimate their potential against international competition. These studies allow for an appreciation of each athlete's dynamics to see their evolution in a measured and objective way.
For confrontation sports, such as wheelchair fencing, para shooting sport, or para archery, rating models are developed to objectively assess the athletes' levels. Specific statistical indicators for each sport are calculated and provided to the teams to optimize their training preparation and competition strategy.
In team sports like wheelchair basketball, individual and collective indicators are calculated to objectify performances and analyze each team's strengths and weaknesses. Many elements are studied, such as the strategy of the level of disability in team composition or the effectiveness of shots from different areas of the court.
The project is involved in 15 different Paralympic sports, allowing for a rich analysis with numerous issues and methodologies set up to support the various actors preparing for the Paris 2024 Paralympic Games.

Keywords: Paralympic Games, High performance, Data analysis

# 1 Introduction

Depuis 2008, l'IRMES (Institut de Recherche bioMédicale et d'Epidémiologie du Sport,) travaille sur la détection des jeunes talents et l'évaluation du potentiel des athlètes. Des variables existent qui permettent d'estimer le potentiel de médailles des athlètes ; cette méthode a été validée dans plus de 50 disciplines. Le parcours des médaillés a montré les particularités du sport paralympique à travers des trajectoires et des durées de carrière variables en fonction du type de handicap (les athlètes en fauteuil roulant maintiennent des performances plus durables par rapport aux athlètes debout). Ces outils de détection ont été complétés par une analyse de la carrière des meilleurs athlètes du monde et une analyse des écarts de performance entre les athlètes préparant Tokyo 2020. Ils sont maintenant utilisés par les entraîneurs pour repositionner les athlètes à haut potentiel dans d'autres épreuves ou disciplines. Cette approche tient compte de la nature du handicap et des niveaux de classification, afin d'estimer les potentiels de performance grâce à des méthodes de suivi spécifiques.

# 2 Analyse sur les sports individuels

Dans le cadre du projet PARAPERF, de nombreux sport avec des métriques tel qu'un chrono ou une distance ont fait l'objet d'étude. C'est le cas du para athlétisme, para natation, para cyclisme et du para haltérophilie. L'un des objectifs principaux du projet est d'étudier la dynamique individuelle des athlètes et de pouvoir estimer leurs potentiels vis-à-vis de la

concurrence. Des modèles permettant d'étudier le lien entre l'âge et la performance chez les valides a été développé dès 1975 par Dan H. Moore qui a modélisé cette relation par une somme de deux exponentielles biphasiques. Puis en 2020, Berthelot et al ont modélisé cette relation par l'équation IMAP (an Integrative Model of Age Performance). Le projet Paraperf a été le premier à mettre en application ces modèles chez les athlètes paralympiques.
Un outil d'aide à **la** décision conçu pour aider les équipes de France ont été développé permettant de suivre les talents ayant une courbe de progression intéressante. Ce qui permet de mettre en lumière certains athlètes avec un certain potentiel même s'ils n'atteignent pas des minima fixe.

**Figure 1 : Trajectoire de performance d'un athlète au 100m T11**

# 3 Analyse sur les sports d'opposition

## 3.1 Mise en place de système de rating

Les sports d'opposition concernés par le projet sont le para tennis de table, le para badminton, la boccia, l'escrime fauteuil, le tir à l'arc et le tir sportif. Il est parfois difficile d'objectiver la performance dans ces sports, car elle dépend directement du niveau des adversaires. Pour cela, des méthodes de ranking ont été utilisées. En reprenant celles proposées dans la littérature scientifique (Elo, Glicko, Stephenson, ...), le projet PARAPERF a mis à disposition des staffs un système de ranking qui permet de classer les adversaires en temps réel, mis à jour à la fin de chaque compétition. Pour chaque modèle, une optimisation des paramètres a été effectuée, et le meilleur modèle (celui qui maximise le taux de bonne prédiction sur l'échantillon test) a été retenu.

**Exemple d'application : Rating en para tennis de table**

Pour quantifier les performances des athlètes par nation nous définissons la fonction $P(a, c)$ représentant les points attribués à un athlète $a$ dans une classification d'handicap $c$ par la formule :

$$P(a,c) = \frac{T(c)}{R(a,c)},$$

où $T(c)$ est le total d'athlètes classés dans la classification $c$ et $R(a,c)$ le rang de l'athlète $a$ dans le classification c. Le score total d'une nation $n$, $S(n)$, est alors donné par

$$S(n) = \sum_{c \in C} \sum_{a \in A_{n,c}} P\,(a,c),$$

avec $A_{n,c}$ l'ensemble des athlètes de $n$ en classification $c$ et C l'ensemble des 22 classifications d'handicap en para tennis de table.

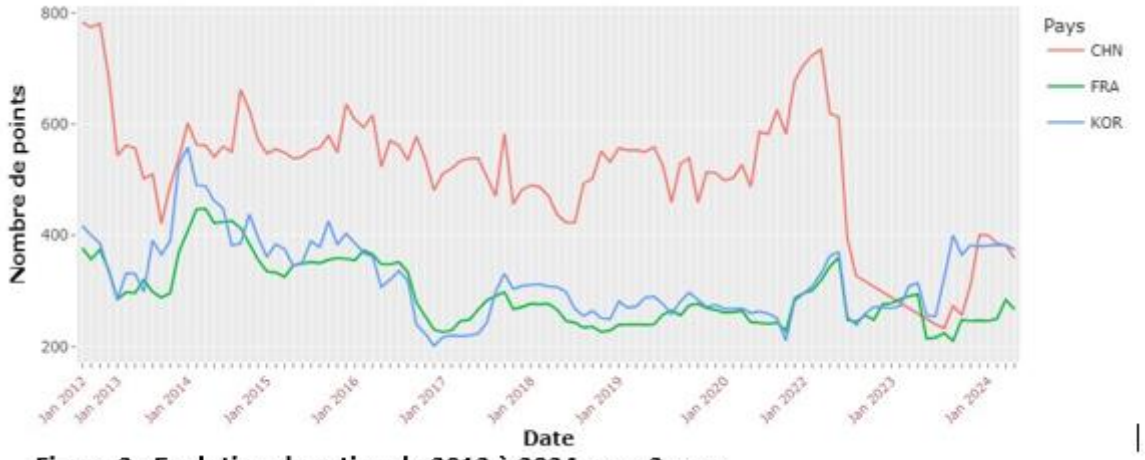


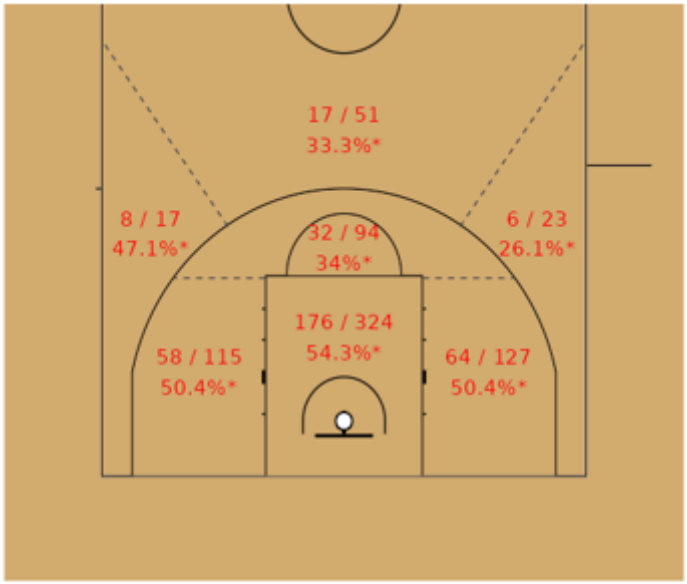

**Figure 2 : Evolution du rating de 2012 à 2024 pour 3 pays**

## 3.2 Mise en place d'autres indicateurs statistiques

Pour mieux caractériser la performance des joueurs, de nombreux indicateurs statistiques ont été calculés. Ils sont propres à chaque sport. Des études sur l'importance du premier set en para tennis de table et en para badminton ont été mises en place, permettant de quantifier de manière individuelle l'importance de ce premier set à l'aide de régression logistique. Pour le tir à l'arc et le tir sportif, qui ont la particularité d'avoir deux phases distinctes : la phase de qualification et la phase de confrontation, le lien entre le classement de la première phase et celui de la deuxième a été étudié, permettant de mieux comprendre l'importance de réaliser une bonne performance dans la première phase. D'autres indicateurs ont été calculés en para tir à l'arc pour mieux comprendre la performance, comme le nombre de « perfects » (lorsque la flèche touche le centre de la cible, cela rapporte le nombre de point maximal), la régularité des archers, d'une part de manière statique mais aussi de manière dynamique au cours de la compétition.

# 4 Analyse sur les sports d’équipes

Dans le cadre des sports d'équipes, tels que le basket fauteuil, PARAPERF met en œuvre une méthodologie détaillée pour l'analyse des performances, combinant des indicateurs individuels et collectifs. Cette approche a pour objectif d'objectiver de manière précise les performances, tout en identifiant les points forts et les points faibles de chaque équipe. Une attention particulière est accordée à divers aspects stratégiques, notamment l'influence du niveau de handicap sur la composition d'équipe et sur la dynamique de jeu. L'efficacité des tirs depuis différentes zones du terrain est également scrutée, permettant d'identifier les secteurs où les joueurs excellent ou, au contraire, où ils rencontrent des difficultés.
Ce processus analytique offre aux entraîneurs des indicateurs pour une compréhension objective des capacités de leurs propres joueurs ainsi que des tactiques employées par les adversaires. En analysant la répartition des handicaps au sein des équipes, les entraîneurs peuvent ajuster leurs stratégies pour exploiter au mieux les points faibles adverses tout en renforçant leurs atouts. La cartographie de l'efficacité des tirs selon les zones du terrain devient un outil puissant pour la préparation tactique, permettant de définir les meilleures positions pour maximiser les chances de marquer tout en identifiant les zones défensives à renforcer.




** : Pourcentage de réussite au tir dans cette zone.*

**Figure 3 : Répartition de la proportion de paniers réussis en basket fauteuil pour l'équipe senior masculine des États-Unis.**

# Bibliographie

Moore, D. H., 2nd. A study of age group track and field records to relate age and running speed. *Nature* **253**, 264–265 (1975).

Berthelot, G. *et al.* An integrative modeling approach to the age-performance relationship in mammals at the cellular scale. *Sci Rep* **9**, 418 (2019).

Le Toquin B, Schipman J, Larochelambert Q, Saulière G, Duncombe S, Toussaint J-F. Is the visual impairment origin a performance factor? Analysis of international-level para swimmers and para athletes. Journal of Sports Sciences (2021).

Omolade, O., & Stephenson, J. Best Rating Scale Design Theory: Implications for Developing Questionnaires in Nursing and Health Sciences. Journal of Modern Nursing Practice and Research (2023).

# Ordonner pour personnaliser : cas d'usage dans le monde du pari sportif

Paul Steffen [1] & Bertrand Beaufils [2]

[1] *Betclic Group, p.steffen@betclicgroup.com*
[2] *Betclic Group, b.beaufils@betclicgroup.com*

**Résumé.** Betclic est un des leaders européens des paris sportifs et des jeux en ligne. Proposer une expérience utilisateur fluide et adaptée à chacun de ses utilisateurs est un défi majeur, notamment pour une application proposant des paris sur plus de 450 évènements sportifs par jour répartis sur près de 50 sports différents. Face à ce challenge, la personnalisation de la page d'accueil de l'application est un enjeu réel, pouvant être solutionné à l'aide d'une modélisation statistique visant à filtrer puis ordonner les évènements sportifs de manière pertinente.
Afin d'étudier la prise de paris de nos utilisateurs sur notre page d'accueil, une base de données de plusieurs milliards d'observations relevées sur 2 ans a été construite. Cette dernière a permis l'entraînement d'un système de recommandation hybride utilisant une factorisation matricielle afin de scorer l'ensemble des éventuelles interactions utilisateur-match, et de proposer un ordonnancement pertinent pour chacun de ces utilisateurs.
Après une présentation de la métrique retenue pour évaluer ce modèle ainsi que du design expérimental choisi pour représenter au mieux la qualité prédictive du modèle une fois mis en production, les alternatives permettant d'éventuellement améliorer cette méthode de classement seront évoquées.

**Mots-clés.** méthode de classement, système de recommandation, pari sportif, factorisation matricielle

**Abstract.** Betclic is one of Europe's leading sports betting and online gaming companies. Providing a smooth and tailored user experience to each of its users is a major challenge, particularly for an application offering bets on more than 450 sporting events a day across almost 50 different sports. To meet this challenge, personalising the application's home page is a real issue, which can be solved with the help of statistical modelling aimed at filtering and ordering sporting events in a relevant way on this page.
In order to study how our users take bets on our home page, a database of several billion observations over 2 years was built and used to train a hybrid recommender system using matrix factorisation to score all possible user-game interactions, and to propose a relevant order for each user.
After a presentation of the metrics used to evaluate this model and the experimental design chosen to best represent the predictive quality of the model once it is in production, we will look at alternatives for improving this ranking method.

**Keywords.** ranking method, recommender system, sports bet, matrix factorization

# 1 La problématique

La personnalisation de la page d'accueil de notre application revient à un problème de classement des différents évènements sportifs disponibles au moment de la consultation de cette page. Ainsi, un premier filtre concernant la date des matchs permet de ne considérer que les évènements non terminés, éligibles au pari sportif. Un second, basé sur une analyse mettant en avant l'intérêt de nos utilisateurs pour les matchs ayant lieu dans un futur proche, a été défini afin de ne conserver que les matchs débutant dans les 7 jours à venir.

# 2 Les données

Avec plus de 2 milliards de paris sportifs uniquement sur les années 2022 et 2023, réalisés par plus de 3,8 millions d'utilisateurs différents, le jeu de données à disposition pour répondre à ce problème est suffisamment important et nécessite même une attention particulière quant aux méthodes utilisées afin de rendre les traitements réalisables sous contrainte de capacité de calcul et de temps.
L'ensemble de ces interactions entre utilisateurs et rencontres sportives (ou *items*) peut être caractérisé à l'aide de différentes variables comme le sport, la compétition et les opposants de la rencontre ou encore la différence temporelle entre le début de la rencontre et le moment où le pari a été effectué. Ainsi, en simplifiant au cas où uniquement 2 opposants se rencontrent lors d'un évènement sportif, on peut obtenir le diagramme de base de données suivant :

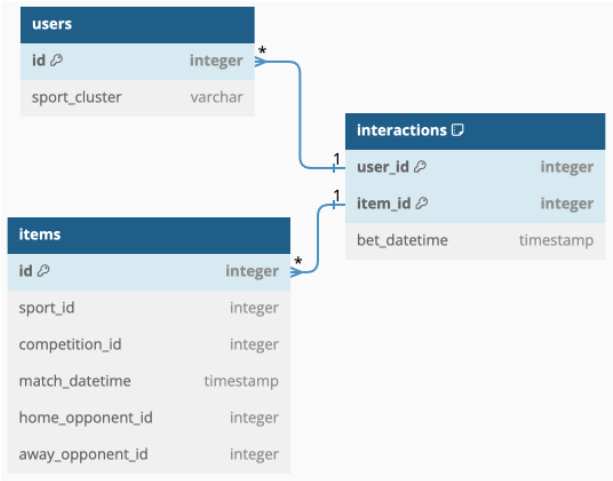



Figure 1: Diagramme de base de données

Nous pouvons alors comprendre que ces données ne représentent que l'observation du comportement de nos utilisateurs. Contrairement aux interactions explicites, provenant d'un retour de la part des utilisateurs et caractérisant à quel point ce dernier a pu apprécier ou non l'objet de la recommandation, nous n'utilisons ici que des interactions implicites. Ne nécessitant aucun retour d'information de l'utilisateur, ce dernier type d'interactions a

l'avantage d'être plus fréquent, bien qu'il ne précise pas l'appréciation ou non de l'utilisateur, en considérant chaque interaction comme positive.
Avec des pics horaires pouvant atteindre plus de 550 000 paris, une quantité importante de données informatives peut être ajoutée en très peu de temps. Les nouveaux utilisateurs s'inscrivant chaque jour et le catalogue de matchs évolutif, imposent que l'ensemble de données utilisé doit régulièrement être mis à jour. Enfin, la particularité d'intégrer régulièrement de nouveaux évènements jusqu'alors inconnus (comme avec le tennis, sport pour lequel il est impossible de connaître les caractéristiques exactes des matchs du tour suivant, sans que les matchs du tour actuel ne soient terminés) et de rendre d'autres matchs inéligibles une fois arrivés à leur terme, implique un contrôle de la fraîcheur des données important.

# 3 Modélisation statistique

## 3.1 Système de recommandation hybride et classement

Face à cette problématique, et avec le jeu de données à notre disposition, un premier choix fut d'utiliser un système de recommandation hybride, combinant le meilleur du filtrage collaboratif et du filtrage basé sur le contenu. Contrairement au premier type de filtrage, utilisant la similarité des utilisateurs basée sur leur historique d'interactions, un système hybride permet d'apporter plus de diversité que ce type de recommandation du fait qu'il ne soit pas limité à recommander des matchs populaires auprès d'utilisateurs similaires. À l'inverse d'un filtrage basé sur le contenu, un système hybride rend possible une certaine sérendipité[1] et évite un effet de bulle de recommandation, en proposant du contenu différent de l'historique d'un utilisateur.
Ainsi, le framework *LightFM* de Kula (2015), implémente un modèle hybride de factorisation de matrices représentant chaque utilisateur et chaque match comme une représentation linéaire de facteurs latents de leurs caractéristiques. Ainsi, pour notre cas, n'ayant pas identifié de caractéristique suffisamment pertinente pour cette problématique, un utilisateur $u \subset U$ n'a été représenté que par un facteur latent, propre à chaque utilisateur : $q_u$ de dimension 30 et un biais $b_u$.
Les matchs $i \subset I$, décrits par des caractéristiques $c_i \subset F^I$ beaucoup plus informatives dans ce problème de recommandation, ont été représentés par un facteur latent composé des *embedding* [2] $e_f^I$, eux aussi de dimension 30 pour chaque caractéristique $c$:

$$p_i = \sum_{j \in c_i} e_j^I$$

et un biais :

$$b_i = \sum_{j \in c_i} b_j^I$$

[1] sérendipité : fait de faire par hasard une découverte inattendue qui s'avère ensuite fructueuse
[2] embeddings: représentation numérique d'une caractéristique

La dimension de ces *embeddings* n'a actuellement pas été ajustée afin d'optimiser la performance du modèle et a été laissée par défaut comme conseillé lors d'une première approche. Ainsi, le score d'une interaction entre un utilisateur $u$ et un match $i$ est obtenu par le produit de leur représentation, ajusté par leurs biais respectifs :

$$\hat{r}_{ui} = f(q_u \cdot p_i + b_u + b_i)$$

où $f(\cdot)$ est la fonction sigmoïde : $f(x) = \frac{1}{1+exp(-x)}$ et l'optimisation du modèle s'effectue par maximum de vraisemblance :

$$L(e^U, e^I, b^U, b^I) = \prod_{(u,i)\in S^+} \hat{r}_{ui} \times \prod_{(u,i)\in S^-} (1 - \hat{r}_{ui})$$

avec $S^+$ et $S^-$ représentant respectivement les interactions positives (paris placés sur le match) et négatives (pari non placé sur un match disponible) entre les utilisateurs et les matchs.
À l'aide de ces représentations latentes des matchs, le problème de démarrage à froid (caractérisant le manque de connaissance d'un nouvel arrivant dans le système de recommandation) des nouveaux matchs du catalogue a ainsi pu être écarté, du fait que le sport, la compétition, et les équipes caractérisant ces matchs soient déjà connus par le modèle.

Avec, en moyenne, plus de 250 000 utilisateurs journaliers différents, proposer un contenu spécifique à chacun d'entre eux représente un réel défi technique pour la fluidité de notre application. C'est pourquoi, les utilisateurs ont été regroupés par une suite de règles déterminées à l'aide de leur historique de paris, réduisant fortement le volume et la complexité de la tâche. Ainsi, il a été possible d'obtenir le classement médian de chacun des matchs, pour un groupe d'utilisateurs similaires, et de proposer un unique classement pour ce même groupe.

## 3.2 Méthode d'évaluation et résultats

Afin d'évaluer la pertinence de notre modèle prédictif, il a été nécessaire de choisir une métrique d'évaluation adaptée à ce problème de classement. Du fait qu'il n'y ait généralement que quelques matchs pertinents pour un utilisateur dans un catalogue disponible pouvant atteindre près de 500 matchs et qu'il est préférable que ces derniers soient mis en avant, le rappel à 5 a été une métrique privilégiée afin de répondre au mieux à cet objectif.

$$Rappel@5 = \frac{nombre\ d'éléments\ pertinents\ dans\ le\ top\ 5}{nombre\ total\ d'éléments\ pertinents}$$

Cependant, le comportement des utilisateurs peut être très varié. Certains parieurs peuvent atteindre un grand nombre d'interactions à l'aide de paris combinés. Ainsi, pour rééchelonner cette métrique et la rendre comparable entre les différents utilisateurs, le dénominateur a été borné à 5. On obtient alors pour notre cas d'usage la formule suivante :

$$Hit@5 = \frac{nombre\ de\ matchs\ pariés\ dans\ le\ top\ 5}{min(nombre\ total\ de\ matchs\ pariés, 5)}$$

Enfin, pour s'assurer du caractère de généralisation du modèle étudié, une validation croisée temporelle a été choisie. Le modèle est alors entraîné avant de faire les prédictions. De cette façon, le modèle utilise toutes les informations disponibles.
Il s'agit d'une variante de la validation croisée standard mais, au lieu de procéder à une distribution aléatoire des observations, l'ensemble d'entraînement est augmenté de manière séquentielle, en conservant l'ordre temporel des données.
Dans l'optique d'optimiser la performance du modèle, la fraîcheur des données joue un rôle essentiel. Ainsi, filtrer l'ensemble des matchs terminés et inclure les nouveaux matchs disponibles toutes les heures, amène à recalculer un classement toutes les heures et permet d'obtenir un résultat au plus proche de son utilisation en production.
En agrégeant les $Hit@5$ sur un ensemble de périodes temporelles de l'année 2023, permettant de représenter au mieux la diversité que peut connaître le calendrier sportif sur une année, un $Hit@5$ d'environ 0,22 a été observé. Cette performance ne peut malheureusement pas être comparée à d'autres cas d'usage similaires dans le secteur d'activité par manque de disponibilité de l'information.

# 4 Conclusion

Bien que ces résultats soient encourageants, ils ne permettent pas à l'heure actuelle d'utiliser un système de recommandation hybride sur notre page d'accueil. Un modèle analytique, calculant un score pour chaque groupe d'utilisateurs basé sur le nombre d'interactions observé par match permet à l'heure actuelle d'obtenir de meilleurs résultats selon la méthodologie d'évaluation précédemment décrite. Un réglage plus fin du système de recommandation laisse penser à de possibles meilleurs résultats dans le futur. Enfin, l'utilisation d'algorithmes de classement tel que *LambdaMART* ou de systèmes de recommandation utilisant des réseaux de neurones comme *DLRM* permettent d'envisager des alternatives face à ce problème d'ordonnancement.

# Bibliographie

Maciej Kula (2015), Metadata Embeddings for User and Item Cold-start recommendations, *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20, 2015.*

Burges, Christopher J. C. (2010), *From RankNet to LambdaRank to LambdaMART: An Overview.*

Naumov, Maxim and Mudigere, Dheevatsa and Shi, Hao-Jun and Huang, Jianyu and Sundaraman, Narayanan and Park, Jongsoo and Wang, Xiaodong and Gupta, Udit and Wu, Carole-Jean and Azzolini, Alisson and Dzhulgakov, Dmytro and Mallevich, Andrey and Cherniavskii, Ilia and Lu, Yinghai and Krishnamoorthi, Raghuraman and Yu, Ansha and Kondratenko, Volodymyr and Pereira, Stephanie and Chen, Xianjie and Smelyanskiy, Misha

(2019), *Deep Learning Recommendation Model for Personalization and Recommendation Systems.*

# Statistique mathématique

# Support and distribution inference from noisy data

Jérémie Capitao-Miniconi[1] & Élisabeth Gassiat [2] & Luc Lehéricy [3]

[1] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France, Jeremie.Capitao-Miniconi@universite-paris-saclay.fr*
[2] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France, Elisabeth.Gassiat@universite-paris-saclay.fr*
[3] *Université Côte d'Azur, CNRS, LJAD, France, Luc.Lehericy@univ-cotedazur.fr*

**Résumé.** Nous considérons des observations bruitées d'une distribution $G$ dont le support est inconnu. Dans le modèle de déconvolution, il a récemment été démontré [6] que, sous des hypothèses très faibles, il est possible de résoudre le problème de déconvolution sans connaître la distribution du bruit et sans aucun échantillon du bruit. Dans ce même modèle, Nous nous intéréssons à estimer la distribution $G$ et son support. Nous commençons par définir les paramètres généraux dans lesquels la théorie s'applique et présentons des classes de supports pouvant être récupérées dans ce contexte. Nous exposons ensuite des classes de distributions pour lesquelles nous prouvons des vitesses minimax adaptatifs (jusqu'à un facteur log log) pour l'estimation du support en distance de Hausdorff. De plus, pour la classe de distributions à support compact, nous fournissons des estimateurs de la distribution inconnue (généralement singulière) et prouvons des vitesses maximums en distance de Wasserstein. Nous démontrons également une borne inférieure presque correspondante sur le risque minimax associé.

**Mots-clés.** Déconvolution, Apprentissage de variété, statistiques non paramétriques, distance de Hausdorff, distance de Wasserstein, bruit inconnu.

**Abstract.** We consider noisy observations of a distribution $G$ with unknown support. In the deconvolution model, it has been proved recently [6] that, under very mild assumptions, it is possible to solve the deconvolution problem without knowing the noise distribution and with no sample of the noise. In this same model, we are interested in estimating the distribution $G$ and its support. We first give general settings where the theory applies and provide classes of supports that can be recovered in this context. We then exhibit classes of distributions over which we prove adaptive minimax rates (up to a log log factor) for the estimation of the support in Hausdorff distance. Moreover, for the class of distributions with compact support, we provide estimators of the unknown (in general singular) distribution and prove maximum rates in Wasserstein distance. We also prove an almost matching lower bound on the associated minimax risk.

**Keywords.** Déconvolution, Manifold learning, Non-paramétric statistics, Hausdorff distance, Wasserstein distance, unknown noise.

# 1 Introduction

It is a common observation that high dimensional data has a low intrinsic dimension. The computational geometry point of view gave rise to a number of interesting algorithms (see [2] and references therein) for the reconstruction of a non linear shape from a point cloud, and in the statistical community, past years have seen increasing interest for manifold estimation. The case of non noisy data, that is when the observations are sampled on the unknown manifold, is by now relatively well understood. When the loss is measured using the Hausdorff distance, minimax rates for manifold estimation are known and have been proved recently. The rates depend on the intrinsic dimension of the manifold and differ when the manifold has a boundary or does not have a boundary, due to the particular way points accumulate near boundaries (see [1] for the most recent results, together with an overview of the subject and references).

When considering the estimation of a distribution with unknown non linear low dimensional support, one has to choose a loss function. The Wasserstein distance allows to compare distributions that can be mutually singular, and is thus useful to compare distributions having possibly different supports. Moreover, approximating an unknown probability distribution $\mu$ by a good estimator $\hat{\mu}$ with respect to the Wasserstein metric allows to infer the topology of the support of $\mu$, see [4]. When using non noisy data, one can look at [5] and [7] for the most recent results and for an overview of the references. However, despite these fruitful developments, geometric inference from noisy data remains a theoretical and practical widely open problem.

In this work, we are interested in the estimation of possibly low dimensional supports, and of distributions supported on such supports, when the observations are corrupted with *unknown* noise. We aim at giving a new contribution on the type of noise which can affect the data without preventing to build consistent estimators of the support and of the law of the noisy signal.

# 2 Setting

We consider independent and identically distributed observations $Y_i$, $i = 1, \dots, n$ coming from the model

$$Y = X + \varepsilon, \tag{1}$$

in which the signal $X$ and the noise $\varepsilon$ are independent random variables. We assume that the observation has dimension at least two, and that its coordinates can be partitioned in such a way that the corresponding blocks of noise variables are independently distributed, that is

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} + \begin{pmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{pmatrix} = X + \varepsilon \tag{2}$$

in which $Y^{(1)}, X^{(1)}, \varepsilon^{(1)} \in \mathbb{R}^{d_1}$ and $Y^{(2)}, X^{(2)}, \varepsilon^{(2)} \in \mathbb{R}^{d_2}$, for $d_1, d_2 \geq 1$ with $d_1 + d_2 = D$, and we assume that the noise components $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ are independent random variables.

We write $G$ the distribution of $X$ and $\mathcal{M}_G$ its support. For $i \in \{1,2\}$, we write $\mathbb{Q}^{(i)}$ the distribution of $\varepsilon^{(i)}$, so that $\mathbb{Q} = \mathbb{Q}^{(1)} \otimes \mathbb{Q}^{(2)}$ is the distribution of $\varepsilon$.

We shall not make any more assumption on the distribution of the noise $\varepsilon$, and we shall not assume that its distribution is known. Indeed in [6], it is proved that under very mild conditions on the distribution of the signal $X$, model (2) is fully identifiable, that is one can recover $G$, and thus its support, and $\mathbb{Q}$ from $G * \mathbb{Q}$.

We introduce the assumptions on the distribution of the signal we shall use. The first one is about the tail of $G$. Let $\rho$ be a positive real number.

**A($\rho$)** There exist $a, b > 0$ such that for all $\lambda \in \mathbb{R}^D$, $\mathbb{E}\left[\exp\left(\lambda^\top X\right)\right] \leq a \exp\left(b\|\lambda\|_2^\rho\right)$.

Under A($\rho$), the characteristic function of the signal can be extended into the multivariate analytic function

$$\begin{aligned} \Phi_X : \mathbb{C}^{d_1} \times \mathbb{C}^{d_2} &\longrightarrow \mathbb{C} \\ (z_1, z_2) &\longmapsto \mathbb{E}\left[\exp\left(iz_1^\top X^{(1)} + iz_2^\top X^{(2)}\right)\right]. \end{aligned}$$

The second assumption is a mild dependence assumption.

**(Adep)** For any $z_0 \in \mathbb{C}^{d_1}$, $z \mapsto \Phi_X(z_0, z)$ is not the null function and for any $z_0 \in \mathbb{C}^{d_2}$, $z \mapsto \Phi_X(z, z_0)$ is not the null function.

The authors of [6] proved the following identifiability Theorem.

**Theorem 1** *if the distribution of the signal satisfies A($\rho$) for some $\rho < 2$ and (Adep), then the distribution of the signal and the distribution of the noise can be recovered from the distribution of the observations up to translation.*

The identifiability result above is the base upon which our estimators are built. A key part of its proof is the following result: if a multi-analytic function $\phi$ satisfies $\phi(0) = 1$, $\phi(t) = \overline{\phi(-t)}$ for all $t$, as well as assumptions A($\rho$) for some $\rho < 2$, (Adep) and

$$\phi(t_1, t_2)\Phi_X(t_1, 0)\Phi_X(0, t_2) = \Phi_X(t_1, t_2)\phi(t_1, 0)\phi(0, t_2)$$

for all $(t_1, t_2)$ in a neighborhood of zero in $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, then $\phi = \Phi_X$. In particular, fixing some $\nu_{\text{est}} > 0$, the only function $\phi$ such that

$$\int_{B^{d_1}_{\nu_{\text{est}}} \times B^{d_2}_{\nu_{\text{est}}}} |\phi(t_1, t_2)\Phi_X(t_1, 0)\Phi_X(0, t_2) - \Phi_X(t_1, t_2)\phi(t_1, 0)\phi(0, t_2)|^2 |\Phi_{\varepsilon^{(1)}}(t_1)\Phi_{\varepsilon^{(2)}}(t_2)|^2 dt_1 dt_2 = 0$$

is $\Phi_X$; the addition of the characteristic functions of the noises $\Phi_{\varepsilon^{(i)}}$, $i \in \{1,2\}$, does not result in any loss of generality since they are continuous and equal to 1 in zero. From there, we construct an empirical criterion $M_n$ that estimates the above integral as, for any multi-analytic function $\phi$,

$$M_n(\phi) = \int_{B^{d_1}_{\nu_{\text{est}}} \times B^{d_2}_{\nu_{\text{est}}}} |\phi(t_1,t_2)\tilde{\phi}_n(t_1,0)\tilde{\phi}_n(0,t_2) - \tilde{\phi}_n(t_1,t_2)\phi(t_1,0)\phi(0,t_2)|^2 dt_1 dt_2, \qquad (3)$$

where for all $(t_1,t_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$,

$$\tilde{\phi}_n(t_1,t_2) = \frac{1}{n}\sum_{\ell=1}^{n} \exp\left\{ it_1^\top Y_\ell^{(1)} + it_2^\top Y_\ell^{(2)} \right\}.$$

We minimize $M_n$ to construct an estimator of $\Phi_X$: let

- $\Upsilon_\rho$ be a set of multivariate analytic functions on $\mathbb{C}^D$, which is in essence the set of functions satisfying A($\rho$) and such that $\phi(0) = 1$ and $\phi(t) = \overline{\phi(-t)}$ for all $t$,
- $\mathcal{H}$ be any closed set of $L^2([-\nu_{\text{est}}, \nu_{\text{est}}]^D)$ of functions satisfying (Adep)
- $\mathbb{C}_m[X]$ be the set of polynomial functions of degree $m$ in $D$ indeterminates. The degree $m$ must satisfy $m \in [2\rho \log n/\log\log n, C\log n/\log\log n]$ for some $C > 2\rho$ arbitrarily large. In the following, we will take $m = \lceil 4\frac{\log n}{\log\log n}\rceil$.

Our estimator of $\Phi_X$ is a $1/n$-minimizer of $M_n$ over the intersection of these three sets:

For any integer $m$ and any $\rho > 1$, let $\widehat{\Phi}_{n,m,\rho}$ be a (up to $1/n$) measurable minimizer of the functional $\phi \mapsto M_n(\phi)$ over $\Upsilon_{\rho,S} \cap \mathcal{H} \cap \mathbb{C}_m[X]$.

Provided $\Phi_X$ itself is in $\Upsilon_\rho \cap \mathcal{H}$, this estimator converges toward $\Phi_X$. As for its rate of convergence, by carefully analyzing the behaviour of $M_n$ near $\Phi_X$. We show the following Proposition.

**Proposition 1 (Variant of Proposition 1 in [3])** *For all $\rho_0 < 2$, $\nu \in (0, \nu_{est}]$, $S, c(\nu), E, C > 0$ and $\delta, \delta', \delta'' \in (0,1)$ with $\delta' > \delta$, there exist positive constants $c$ and $n_0$ such that the following holds: let $\rho \in [1, \rho_0]$, for all $\Phi_X \in \Upsilon_{\rho,S} \cap \mathcal{H}$ and $\mathbb{Q} \in \mathcal{Q}^{(D)}(\nu, c(\nu), E)$, for all $n \geq n_0$ and $x \in [1, n^{1-\delta'}]$, with probability at least $1 - 2e^{-x}$,*

$$\sup_{\rho' \in [\rho,\rho_0],\ m \in [2\rho' \frac{\log n}{\log\log n}, C\frac{\log n}{\log\log n}]} \int_{B^{d_1}_\nu \times B^{d_2}_\nu} |\widehat{\Phi}_{n,m,\rho'}(t) - \Phi_X(t)|^2 dt \leq c\left(\frac{x}{n^{1-\delta}}\right)^{1-\delta''}.$$

# 3 Estimation

From there, we first consider the estimation of its support $\mathcal{M}_G$. While it may not be possible to directly recover the distribution of $X$ from $\Phi_X$ by inverse Fourier transform, since $G$ could be singular, it is possible to recover the function $G * \Psi$ for any properly chosen kernel $\Psi$.

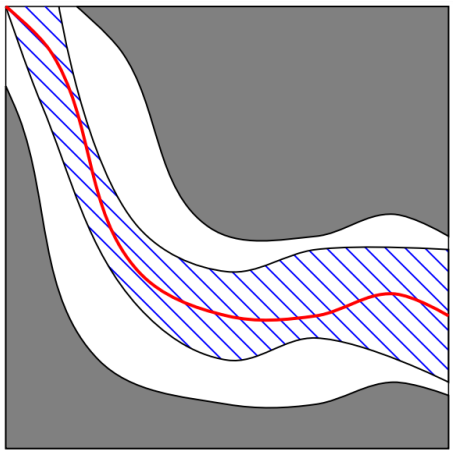


Figure 1: In red, the support of the signal distribution $\mathcal{M}_G$, the blue hatched area represents the set $\{\bar{g} \geq \lambda_n + \|\bar{g} - \widehat{g}_n\|_\infty\}$ and the gray area represents the set $\{\bar{g} \leq \lambda_n - \|\bar{g} - \widehat{g}_n\|_\infty\}$, so that the estimator of the support lies in between the gray and the blue areas.

Moreover, when $\Psi$ is an approximation of the Dirac, the support of $G$ can be approximated by an upper level set of $G * \Psi$. Intuitively, if $\Psi$ were the density of a random variable $Z$, $G * \Psi$ would be the density of $X + Z$. When $\Psi$ is an approximation of the Dirac, the distribution of $Z$ will be close to the Dirac in zero, thus $X + Z$ will concentrate close to the support of $X$. Figure 1 illustrates this idea.

The only geometric assumption we need on $G$ is that it is $(a, d)$-standard, *i.e.* that the weight of any ball of radius $r$ centered on a point of its support is at least $ar^d$. In particular, we do not require that the reach of the support $\mathcal{M}_G$ is lower bounded, or even that $\mathcal{M}_G$ is a manifold.

The precise choice of $\Psi$ is non trivial and will not be discussed here. For $f$ an integrable function from $\mathbb{R}^D$ to $\mathbb{R}$, we denote by $\mathcal{F}[f]$ (resp. $\mathcal{F}^{-1}[f]$) the (resp. inverse) Fourier transform of $f$ defined, for all $y \in \mathbb{R}^D$, by

$$\mathcal{F}[f](y) = \int e^{it^\top y} f(t) dt \quad \text{and} \quad \mathcal{F}^{-1}[f](y) = \left(\frac{1}{2\pi}\right)^D \int e^{-it^\top y} f(t) dt.$$

Among the properties of $\Psi$, one is that its Fourier transform $\mathcal{F}[\Psi]$ is compactly supported. We show that the estimator

$$\widehat{g} := \mathcal{F}^{-1}\left[\widehat{\phi} \cdot \mathcal{F}[\Psi]\right]$$

is close to $\bar{g} := \mathcal{F}^{-1}\left[\Phi_X \cdot \mathcal{F}[\Psi]\right] = G * \Psi$, whose upper level sets are close to $\mathcal{M}_G$. Our estimator is thus taken as

$$\widehat{\mathcal{M}} = \{y : \widehat{g}(y) > \lambda\},$$

where $\lambda$ has to be choosen. We also show that, when $H$ is the Hausdorff distance and $\mathcal{K}$ is a known compact in $\mathbb{R}^D$,

$$\mathbb{E}[H(\mathcal{M}_G \cap \mathcal{K}, \widehat{\mathcal{M}} \cap \mathcal{K})] = O\left(\frac{(\log\log n)^{1/\rho+1+\delta}}{(\log n)^{1/\rho}}\right)$$

for some $\delta > 0$ arbitrarily small (that depends on $\Psi$). A caveat of the above result is that we can only control what happens inside a compact set $\mathcal{K}$. On the other hand, $\mathcal{K}$ can be taken arbitrarily large.

We use the two-point method to prove that this result is quasi-optimal, in the sense that for any estimator $\widehat{\mathcal{M}}$, there exists $G$ such that the error $\mathbb{E}[H(\mathcal{M}_G \cap \mathcal{K}, \widehat{\mathcal{M}} \cap \mathcal{K})]$ is at least $\frac{1}{(\log n)^{1/\rho}}$ when $\rho > 1$, and $\frac{1}{(\log n)^{1+\delta}}$ when $\rho = 1$ (i.e. when the support of $X$ is compact), for $\delta > 0$ arbitrarily small.

All these results can be made adaptive in the tail parameter $\rho$; we provide a method based on Goldenshluger and Lepski's method.

We consider the estimation of $G$ in Wasserstein distance, when $G$ is compactly supported. We do so by restricting $\max(\widehat{g}, 0)$ to the estimated support $\widehat{\mathcal{M}}$, then renormalizing it to obtain the probability density of a measure $\widehat{P}$, which satisfies

$$\mathbb{E}[W_2(G, \widehat{P})] = O\left(\frac{\log\log n}{\log n}\right),$$

Again, we use the two-point method to prove that this result is quasi-optimal, showing that for any estimator $\widehat{P}$, there exists $G$ such that the error $\mathbb{E}[W_2(G, \widehat{P})]$ is at least $\frac{1}{(\log n)^{1+\delta}}$, for $\delta > 0$ arbitrarily small

# References


[1] Eddie Aamari, Catherine Aaron, and Clément Levrard. *Minimax Boundary Estimation and Estimation with Boundary*. 2021. DOI: 10.48550/ARXIV.2108.03135. URL: https://arxiv.org/abs/2108.03135.

[2] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and topological inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2018, pp. xii+233. ISBN: 978-1-108-41089-2; 978-1-108-41939-0. DOI: 10.1017/9781108297806. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1017/9781108297806.

[3] Jérémie Capitao-Miniconi and Élisabeth Gassiat. "Deconvolution of spherical data corrupted with unknown noise". In: *Electron. J. Stat.* (to appear).

[4] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. "Geometric inference for probability measures". In: *Found. Comput. Math.* 11.6 (2011), pp. 733–751. ISSN: 1615-3375. DOI: 10.1007/s10208-011-9098-0. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1007/s10208-011-9098-0.

[5] Vincent Divol. "Measure estimation on manifolds: an optimal transport approach". In: *Probab. Theory Related Fields* 183.1-2 (2022), pp. 581–647. ISSN: 0178-8051. DOI: 10.1007/s00440-022-01118-z. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1007/s00440-022-01118-z.

[6] Élisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. “Deconvolution with unknown noise distribution is possible for multivariate signals”. In: *Ann. Statist.* 50.1 (2022), pp. 303–323.

[7] Jonathan Niles-Weed and Quentin Berthet. “Minimax estimation of smooth densities in Wasserstein distance”. In: *Ann. Statist.* 50.3 (2022), pp. 1519–1540. ISSN: 0090-5364. DOI: 10.1214/21-aos2161. URL: https://doi-org.ezproxy.universite-paris-saclay.fr/10.1214/21-aos2161.

# Benign overfitting et régression non-paramétrique adaptative

Julien Chhor[1], Suzanne Sigalla[2] and Alexandre B. Tsybakov[3]

[1] *Toulouse School of Economics, France, julien.chhor@tse-fr.eu*
[2] *CREST/ENSAE, France, suzanne.sigalla@ensae.fr*
[3] *CREST/ENSAE, France, alexandre.tsybakov@ensae.fr*

**Résumé.** Le benign overfitting est un phénomène contre-intuitif récemment découvert dans le cadre du deep learning. Il a été observé expérimentalement que les réseaux de neurones profonds peuvent dans certains cas parfaitement overfitter des données d'entraînement bruitées, tout en ayant d'excellentes performances de généralisation pour prédire de nouveaux points de données. Cela va à l'encontre du point de vue statistique conventionnel selon lequel il devrait y avoir un compromis nécessaire entre le biais et la variance. Ce papier vise à comprendre le benign overfitting dans le cadre simplifié de la régression non paramétrique. Nous proposons d'utiliser des polynômes locaux pour construire un estimateur de la fonction de régression avec les deux propriétés suivantes. Premièrement, cet estimateur est optimal au sens minimax sur les classes de Hölder. Deuxièmement, il s'agit d'une fonction continue qui interpole l'ensemble des observations avec une grande probabilité. L'élément clé de la construction est l'utilisation de noyaux singuliers. De plus, nous démontrons que l'adaptation à une régularité inconnue est compatible avec le surajustement bénin : nous proposons en effet un autre estimateur interpolant qui atteint l'optimalité minimax de manière adaptative à la régularité de Hölder inconnue. Nos résultats mettent en lumière le fait que, dans le modèle de régression non paramétrique, l'interpolation peut être fondamentalement découplée du compromis biais-variance.

**Mots-clés.** Régression non-paramétrique, adaptation, benign overfitting, estimateurs par polynômes locaux, agrégation.

**Abstract.** Benign overfitting is a counter-intuitive phenomenon recently discovered in the context of deep learning. It has been experimentally observed that in certain cases, deep neural networks can perfectly overfit noisy training data, while still achieving excellent generalization performance for predicting new data points. This goes against the conventional statistical viewpoint which posits that there should be a necessary tradeoff between bias and variance. This paper aims to understand benign overfitting in the simplified setting of nonparametric regression. We propose using local polynomials to construct an estimator of the regression function with the following two properties. First, this estimator is minimax-optimal over Hölder classes. Second, it is a continuous function that interpolates the set of observations with high probability. The key element of the construction is the use of singular kernels. Moreover, we demonstrate that adaptation to unknown smoothness is compatible with benign overfitting: indeed, we propose another interpolating estimator that achieves minimax optimality adaptively to the unknown Hölder smoothness. Our results highlight that in the nonparametric regression model, interpolation can be fundamentally decoupled from the bias-variance tradeoff.

**Keywords.** Nonparametric regression, adaptation, benign overfitting, local polynomial estimators, aggregation.

# 1 Texte long

Benign overfitting is a counter-intuitive phenomenon recently discovered in the deep-learning community. Empirically, deep neural networks can conciliate two seemingly conflicting properties: 1) perfect *overfitting* (namely, zero error when predicting a data point in the training set) and 2) excellent prediction accuracy outside of the training data Zhang et al. (2021). Overfitting was previously believed to deteriorate statistical performance. Given neural networks' ubiquity, identifying contexts where the phenomenon of benign overfitting can occur is of practical interest. In this paper (Chhor et al., 2024), we study this phenomenon in the simplified setting of non-parametric regression, defined below.

**State of the art:** Benign overfitting has not yet been explained in the context of deep neural networks, and it was mostly studied in simplified statistical settings: linear regression Bartlett et al. (2020), kernel regression Liang et al. (2020) and more recently, in non-parametric regression Belkin et al. (2019). The only paper examining benign overfitting with non-asymptotic guarantees was Belkin et al. (2019), considering only the case $\beta \in (0, 1]$ and $\beta \in (1, 2]$ under additional smoothness constraints on the density of the design.

**Model and contributions:** For $n, d \geq 1$, assume that we observe i.i.d. random pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \dots, n$ such that

$$\forall i \in \{1, \dots, n\} : Y_i = f(X_i) + \xi_i$$

where the unknown link function $f$ to estimate belongs to a suitably defined Hölder class of functions with smoothness $\beta > 0$ and Hölder constant $L > 0$, denoted as $\Sigma(\beta, L)$. The design random variables $X_i$'s are i.i.d. with density $p$ that has a convex compact support and such that $0 < c \leq p(\cdot) \leq C < \infty$ over the support of $p$ for some $c, C > 0$, and the noise random variables $(\xi_i)_i$ are i.i.d. with a finite moment of order $2 + \delta$ for some $\delta > 0$.

The goal is to construct an estimator $\widehat{f}$ of $f$ that interpolates the training dataset $(X_i, Y_i)_i$ with high probability (i.e. satisfying $\forall i \in \{1, \dots, n\} : \hat{f}(X_i) = Y_i$), while being minimax optimal for the quadratic risk over the Hölder class with smoothness $\beta > 0$:

$$\sup_{f \in \Sigma(\beta, L)} \mathbb{E}\left[\|\hat{f}(X) - f(X)\|^2_{L^2(X)}\right] \leq n^{-2\beta/(2\beta+d)}.$$

We propose a local polynomial estimator of order $\beta$ that achieves the two requirements, thereby showing that benign overfitting is compatible with minimax optimality over the whole scale of Hölder classes. The key element of the construction is the use of a singular kernel.

Then, we use an aggregation technique to obtain a second estimator attaining this optimal rate adaptively to the unknown smoothness $\beta > 0$.

Along the way, we strengthen the theory of local polynomial estimators, by proving that they can produce optimal estimators under much milder assumptions than previously known conditions.

Our results highlight that in nonparametric regression, interpolation can be fundamentally decoupled from the bias-variance tradeoff.

# References


Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2019). Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR.

Chhor, J., Sigalla, S., and Tsybakov, A. B. (2024). Benign overfitting and adaptive nonparametric regression. *Probability Theory and Related Fields (to appear); arXiv preprint arXiv:2206.13347*.

Liang, T., Rakhlin, A., and Zhai, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

# Random coefficients regressions and radon transform inversion using mollification

Pierre Maréchal [1] & Anne Vanhems [2]

[1] *Mathematical Institute of Toulouse, Paul Sabatier University, France, pr.marechal@gmail.com*
[2] *TBS Education, France, a.vanhems@tbs-education.fr*

**Résumé.** Cet article considère le modèle de coefficients aléatoires et propose une approche par mollification pour régulariser la mal-position induite par l'inversion de la transformée de Radon. Nous analysons les propriétés asymptotiques et en déduisons les taux de convergence. Nous comparons notre estimateur aux méthodes classiques existantes à l'aide de simulations.

**Mots-clés.** Modèle à coefficients aléatoires, Problèmes inverses, Transformée de Radon, Mollification

**Abstract.** This paper considers the random coefficients model setting and proposes a mollification approach to regularize the ill-posedness induced by the inversion of the radon transform. We analyse asymptotic properties and derive rates of convergence. We compare our estimator with classical existing methods using simulations.

**Keywords.** Random coefficient models, inverse problems, radon transform, mollification

# 1 Model

Linearity in a causal relationship between a dependent variable and a set of regressors is a common assumption throughout economics. In this paper we consider the case when the coefficients in this relationship are random and distributed independently from the regressors. Our aim is to identify and estimate the distribution of the coefficients nonparametrically. The analysis of such model involves the inversion of the Radon transform, which is an ill-posed inverse problem to solve, see Beran and Hall (1992), Beran et al. (1996), Holderlein et al. (2010) for a presentation of this model.

Following Holderlein et al. (2010), we consider the following regression model

$$Y_i = b_i^T X_i,$$

where $Y_i$ is an observed continuously distributed random scalar, $X_i$ denotes an observed random $d$-vector of individual specific regressors and $b_i$ is an unobserved random $d$-vector of individual coefficients. We assume that $Y \in \mathscr{R}$, $X \in \mathscr{R}^d$ and $b$ is independent from $X$.

We first consider the following transformation of the random variables: $(Y_i, X_i) \mapsto (U_i, S_i)$, $i = 1, ..., n$ where

$$S_i = \|X_i\|^{-1} X_i \in \mathscr{S}_{d-1}, \quad U_i = \|X_i\|^{-1} Y_i \in \mathscr{R},$$

and the unit sphere in $\mathscr{R}^d$ is denoted by $\mathscr{S}_{d-1} = \{z \in \mathscr{R}^d : \|z\| = 1\}$. Then, the model becomes:

$$U_i = b_i^T S_u.$$

Second, we consider the $d-1$ hyperplanes defined by a direction vector $s \in \mathscr{S}_{d-1}$ and a distance from the origin $u \in \mathscr{R}$:

$$P_{s,u} = \{z \in \mathscr{R}^d : \ z^T s = u\}.$$

The Radon transform of a function $f : \mathscr{R}^d \to \mathscr{R}$ is defined as

$$(Rf)(s,u) = \int_{P_{s,u}} f.$$

The conditional density of $U$ given $S$ is given by the Radon transform of the density $f_b$:

$$f_{U|S}(u|s) = Rf_b(s,u).$$

To solve this ill-posed inverse problem and recover the function $f_b$, Holderlein et al. (2010) propose to use a regularized inverse $A_h$ of the Radon transform:

$$(A_h g)(z) = \int_{S_{d-1}} \int_{-\infty}^{\infty} K_h(s^T z - u) g(s,u) du d\mu(s),$$

where $\mu$ is the Lebesgue measure on $S_{d-1}$ and $K$ is a smoothing kernel. A regularized solution for the density $f_b$ could then be defined as:

$$f_b(z) = (A_h f_{U|S)}(z) = \int_{S_{d-1}} \int_{-\infty}^{\infty} K_h(s^T z - u) f_{U|S}(u|s) du d\mu(s).$$

# 2 Mollification

In this paper, we propose a different method to regularize the deconvolution problem, which uses a regularization principle introduced in the deterministic setting, and has been applied in several fields of signal and image processing such as deconvolution of images in astronomy, computerized tomography, as in as in Alibaud et al. (2009). We refer to it as the regularization by mollification, or merely as the mollification, Bonnefond and Maréchal (2009) for a general presentation of the method.

For general ill-posed equations $Rf_b = f_{U|S}$, the variational setting consists in defining a solution *via* an optimization problem the form

$$\text{Minimize} \quad \mathscr{D}(Rf_b, f_{U|S}) + \mathscr{R}(f_b).$$

Here, the first term (referred to as the *data fidelity term*) invites the solution to somehow respect the model while the second term (referred to as the *regularization term*) aims at stabilizing the solution with respect to perturbations of the data $f_{U|S}$.

We call *mollification* the variational counterpart of approximate inverses. The same target, namely

$$C_\beta f_b := \varphi_\beta * f_b,$$

is sought for, but we now use the variational setting. From obvious heuristics, the undesired component of $f_b$ is $(I - C_\beta)f_b$, which suggests that the regularization term should merely be $\mathscr{R}(f_b) := \|(I - C_\beta)f_b\|^2$. As for the data fidelity term, we observe that the target $\varphi_\beta * f_b^\dagger$ satisfies the equation

$$R(\varphi_\beta * f_b) = \varphi_\beta * f_{U|S} = C_\beta f_{U|S},$$

since the radon transform possesses similar associativity and commutativity properties as the convolution. This suggests to take a data fidelity term of the form

$$\mathscr{D}(Rf, f_{U|S}) = \left\|Rf - C_\beta f_{U|S}\right\|^2.$$

In summary, the original mollification approach to the radon transform problem consists in defining the reconstructed density as the (unique) solution to the minimization problem

$$f_{\mathrm{MO},\beta} := \mathrm{argmin}_f \left[\left\|C_\beta f_{U|S} - Rf\right\|^2 + \left\|(I - C_\beta)f\right\|^2\right]. \tag{1}$$

We will also consider the modified version of the mollification, as in Hohage et al. (2022):

$$f_{\mathrm{MM},\beta} := \mathrm{argmin}_f \left[\left\|f_{U|S} - Rf\right\|^2 + \left\|(I - C_\beta)f\right\|^2\right].$$

By the first order optimality conditions we have the explicit formula

$$f_{\mathrm{MO},\beta} = \left(R^*R + (I - C_\beta)^*(I - C_\beta)\right)^{-1} R^* C_\beta f_{U|S}.$$

We note that $R^*R$ is positive definite since $R$ is injective, so that the operator $R^*R + (I - C_\beta)^*(I - C_\beta)$ is also positive definite.

Using mollification approach, we build a regularized estimator of the radon transform inversion that is proved to be convergence with optimal rates of convergence. We also compare the finite sample properties of our estimator with other estimators found in the literature.

# Bibliographie

Alibaud, N., Maréchal, P., and Saesor, Y. (2009). A variational approach to the inversion of truncated Fourier operators. Inverse Problems, 25(4), 045002.

Beran, R., and Hall, P. (1992). Estimating coefficient distributions in random coefficient regressions. The annals of Statistics, pp. 1970-1984.

Beran, R., Feuerverger, A., and Hall, P. (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. The Annals of Statistics, 24(6), pp. 2569-2592.

Bonnefond, X., and Maréchal, P. (2009). A variational approach to the inversion of some compact operators. Pacific journal of optimization, 5(1), pp. 97-110.

Hoderlein, S., Klemela, J., and Mammen, E. (2010). Analyzing the random coefficient model nonparametrically. Econometric Theory, 26(3), pp. 804-837.

Hohage, T., Maréchal, P., Simar, L., and Vanhems, A. (2022). A mollifier approach to the deconvolution of probability densities. Econometric Theory, pp. 1-40.

# Tests Convergents, "Distribution-Free" et Affine-Invariants des Hypothèses du Modèle à Composantes Indépendantes

Marc Hallin [*1] & Simos Meintanis [2] & Klaus Nordhausen [3]

[1] *Université libre de Bruxelles, Belgique, mhallin@ulb.be*
[2] *Université Nationale et Capodistrienne d'Athènes, Grèce, simosmei@econ.uoa.gr*
[3] *Université de Jyväskylä, Finlande, klaus.k.nordhausen@jyu.fi*

**Résumé.** Nous proposons une famille de tests de la validité des hypothèses sous-jacentes aux méthodes de l'analyse en composantes indépendantes. Ces tests reposent sur des statistiques de type $L_2$ pondérées et font intervenir les fonctions caractéristiques empiriques des composantes indépendantes estimées; pour un choix approprié des pondérations et de la méthode d'estimation des matrices de *"demixing"* ces tests sont convergents et invariants par transformation affine. La loi asymptotique sous l'hypothèse de la statistique de test est cependant trop complexe pour une implémentation pratique, et nous établissons la validité en échantillons finis de valeurs critiques permutationnelles ou de rééchantillonnage tirant parti des symétries du problème (en dimension $p$, un groupe de permutations de $(n!)^{p-1}$ éleéments). Ces dernières conduisent à des tests "distribution-free" en dépit du fait que permutations et rééchantillonnage sont effectués sur les composantes indépendantes *estimées* ou leurs rangs marginaux. Les simulations établissent les bonnes performances de ces tests et leur supériorité par rapport à leur unique compétiteur, basé sur la notion de *"distance covariance".*

**Mots-clés.** Fonctions caractéristique; indépendence totale; composantes indépendantes; tests de rangs.

**Abstract.** We propose a family of tests of the validity of the assumptions underlying independent component analysis methods. The tests are formulated as L2–type procedures based on characteristic functions and involve weights; a proper choice of these weights and the estimation method for the mixing matrix yields consistent and affine-invariant tests. Due to the complexity of the asymptotic null distribution of the resulting test statistics, implementation is based on permutational and resampling strategies. This leads to distribution-free procedures regardless of whether these procedures are performed on the estimated independent components themselves or the componentwise ranks of their components. A Monte Carlo study involving various estimation methods for the mixing matrix, various weights, and a competing test based on distance covariance is conducted under the null hypothesis as well as under alternatives.

**Keywords.** Characteristic function; total independence; independent component model; rank test.

*Marc Hallin gratefully acknowledges the support of the Czech Science Foundation grants GAČR22036365 and GA24-10078S.

# 1 Testing the validity of the independent component model

Consider the *independent component model* (ICM)

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Omega}\boldsymbol{Z} \tag{1}$$

whereby the $p$-dimensional random vector $\boldsymbol{X} \coloneqq (X_1, ..., X_p)^\top$, $p > 1$ is linearly associated with a latent random vector $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega}) \coloneqq (Z_1(\boldsymbol{\mu}, \boldsymbol{\Omega}), ..., Z_p(\boldsymbol{\mu}, \boldsymbol{\Omega}))^\top = (Z_1, ..., Z_p)^\top$ of centered independent components, i.e., $\boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ is enjoying the property of *total independence*:[1] the $p$-dimensional distribution of $\boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ is the product of the (marginal) distributions of its components $Z_1(\boldsymbol{\mu}, \boldsymbol{\Omega}), \ldots, Z_p(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

Denoting by $\varphi_\ell$, $\ell = 1, ..., p$ and $\varphi$ the characteristic functions (CFs) of $Z_\ell(\boldsymbol{\mu}, \boldsymbol{\Omega})$ and $\boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, respectively, the property of total independence is equivalent to

$$\varphi(\boldsymbol{t}) = \prod_{\ell=1}^{p} \varphi_\ell(t_\ell), \qquad \boldsymbol{t} = (t_1, ..., t_p)^\top \in \mathbb{R}^p. \tag{2}$$

The linear structure of (1) involves a mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and a $(p \times p)$ matrix $\boldsymbol{\Omega}$, often termed the *mixing matrix*, which belongs to the space $\mathbb{M}^p$ of full-rank $(p \times p)$ matrices.

The independent component model (ICM) is underlying the broad class of methods known as *independent component analysis* (ICA). A widely used method—a Google search (google.com accessed on 05.03.2024) returns no less than 532,000,000 links!—ICA originally was developed in the engineering literature on signal processing, where it has countless applications. It recently also attracted much interest in finance and economics: see, for instance, Comon and Jutten [2010], Garcia-Ferrer et al. [2012], Matteson and Tsay [2017], Gourieroux et al. [2017], Hai [2020], or Miettinen et al. [2020], to quote only a few. All these applications crucially rely on the validity of the ICM assumptions (1)–(2) in some observed sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of i.i.d. copies of $\boldsymbol{X}$. Somewhat surprisingly, though, no test of the validity of these assumptions, i.e., of the the null hypothesis

$$\mathcal{H}_0\colon \text{ the random vector } \boldsymbol{X} \text{ is such that (1) and (2) hold for some } (\boldsymbol{\mu}, \boldsymbol{\Omega}) \in \mathbb{R}^p \times \mathbb{M}^p \tag{3}$$

is available in the literature except for a procedure proposed in Matteson and Tsay [2017] who are using the notion of *distance covariance* and resampling-based critical values. Their test statistic, however, depends on the arbitrary ordering of the components of $\boldsymbol{Z}$, which is not a desirable property. Our objective here is to develop tests of $\mathcal{H}_0$ which, as we shall see, are universally consistent and outperform the Matteson and Tsay procedure.

Our tests are based on CFs and the fact that $\mathcal{H}_0$ holds if and only if there exists $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ such that (2) is satisfied. More precisely, our test statistics are weighted measures of the discrepancy between the (empirical) joint CF and the product of marginal ones—a classical and quite natural measure of total dependence that can be traced back, e.g., to Csörgő and

[1] While *total independence* (2) implies *pairwise independence*—namely, $Z_j \perp\!\!\!\perp Z_k$ for all $j \neq k = 1, \ldots, p$ where $\perp\!\!\!\perp$, as usual, denotes stochastic independence—the converse, of course, is not true.

Hall [1982] and is also considered in Chen and Bickel [2005] in an estimation context. Nobody ever used these measures of total dependence as test statistics nor provided their asymptotic distributions under the assumption of total independence. The reason for this is that these asymptotic distributions (the distributions of infinite sums of chi-squared random variables weighted by the eigenvalues of a complicated integral operator) are hopelessly untractable: indeed, they not only depend on the actual distribution of the observations but also on the way (FOBI, JADE, FastICA, ... ) the mixing matrix is estimated. They cannot be used in the construction of asymptotic critical values and hence are entirely useless in practice. This is probably the reason why these very natural characteristic-function-based statistics so far have not been used as test statistics.

The main and nontrivial theoretical novelty of this contribution is to show that these impracticable asymptotic critical values can be dispensed with and replaced with exact permutational ones. These permutational critical values, however, follow from a non-standard permutational scheme—not the usual $n!$ permutations of the observations, which leave the test statistics invariant! Moreover, the validity of our permutational approach is anything but straightforward, since it involves permutations (in dimension $p$, there are $(n!)^{p-1}$ of them) of the components of estimated residuals that are not i.i.d. under $\mathcal{H}_0$.

Our tests, unlike the Matteson and Tsay ones, are affine-invariant and have exact size $\alpha$ uniformly over $\mathcal{H}_0$, irrespective of the actual (absolutely continuous) distribution of the observations and the demixing method (FOBI, JADE, FastICA, ... ) adopted for deriving the residuals $\widehat{\mathbf{Z}}$. They involve weights $W$; when choosing them, we pay special attention to their impact on the computation of the corresponding test statistics, a feature that is particularly important in case of high-dimensional ICMs.

Now, beyond its ICM structure, $\mathcal{H}_0$ does not specify anything about $\boldsymbol{X}$; in particular, the distribution of $\boldsymbol{Z}$, hence that of $\boldsymbol{X}$, remains completely unspecified under $\mathcal{H}_0$ as well as under the alternative (which includes arbitrary dependencies between the components of $\boldsymbol{X}$). In principle, that distribution even could be discrete, although we are focusing on the absolutely continuous case for simplicity. The problem, therefore, is of a semiparametric nature, where the nuisance parameters are unspecified distributions. Ranks, in that context, naturally come into the picture, and we also propose rank-based tests of $\mathcal{H}_0$ which for the same reason as above, admit fully distribution-free exact critical values.

While the problem of testing bivariate independence has been treated quite extensively in the literature, the results on testing total independence are less abundant; see, however, Blum et al. [1961], Deheuvels [1981], Csörgő [1985], Kankainen [1995]. The related problem of testing vector independence—independence between vectors of finite dimension—recently attracted renewed interest: we refer to Roy et al. [2020], Genest et al. [2019], Herwartz and Maxand [2020], Shi et al. [2022a] for reviews of the existing literature. CF-based methods in this context date back to Csörgő and Hall [1982], Csörgő [1985], Feuerverger [1993], Kankainen and Ushakov [1998]; for more recent contributions, see Meintanis and Iliopoulos [2008], Fan et al. [2017], Pfister et al. [2017], Chakraborty and Zhang [2019], among others. In all these tests, an empirical contrast between the joint and the product of marginal CFs is considered. The advantages, in this context, of the CF-based approach over competing methods based, e.g., on distribution function are confirmed by the success of distance covariance methods

and their relation [Sejdinovic et al., 2013] to RKHS statistics; see Edelmann et al. [2019], Székely and Rizzo [2023], and Chen et al. [2019] for reviews on distance covariance and related approaches, and Eriksson and Koivunen [2003] and Chen and Bickel [2005] for a discussion of their advantages in the ICM context.

As for the advantages of rank-based procedures in the presence of unspecified distributions, we refer to Shi et al. [2022b], who show that pseudo-Gaussian methods such as Wilks' classical test for vector independence [Wilks, 1935], although asymptotically valid, severely over-reject under skewed distributions. Rank-based independence testing methods therefore are a natural solution and have been thoroughly investigated in the bivariate context—see, e.g., Chapters II.4.11 and III.6 of Hájek and Šidák [1967]. With the recent introduction of measure transportation-based concepts of multivariate ranks and signs [Chernozhukov et al., 2017, Hallin et al., 2021], this approach to bivariate independence testing has been extended to arbitrary dimensions [Shi et al., 2022a,b, 2023, Ghosal and Sen, 2022]. The problem we are facing here, again, is trickier. Independence, in these references, is to be tested between observable quantities which under the null are i.i.d. so that the distribution-freeness of rank-based statistics is straightforward while we are dealing with unobservable variables $Z_{j\ell}$, the values of which have to be estimated as $\widehat{Z}_{j\ell}$. These "estimated residuals" are no longer i.i.d. under the null, and handling them requires additional care.

Still in the ICM context, rank-based tests of vector independence have been proposed by Oja et al. [2016]. Their problem is quite different from ours, though: instead of the validity of the ICM assumptions, these authors are testing the independence of two given subvectors of $\boldsymbol{X}$ under maintained ICM assumptions.

## 2 The test statistics

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, with $\boldsymbol{X}_j \coloneqq (X_{j1}, \ldots, X_{jp})^\top$, $j = 1, ..., n$ denote a sample of $n$ independent copies of $\boldsymbol{X}$. To implement the test, since we only observe $\boldsymbol{X}_j$ and not $\boldsymbol{Z}_j$, the latent variables $(\boldsymbol{Z}_j,\ j = 1, ..., n)$ first need to be estimated from the data. Specifically, our test statistic is a function of the "estimated ICM residuals"

$$\widehat{\boldsymbol{Z}}_j = \widehat{\boldsymbol{Z}}_j(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Omega}}_n) \coloneqq \widehat{\boldsymbol{\Omega}}_n^{-1}\left(\boldsymbol{X}_j - \widehat{\boldsymbol{\mu}}_n\right) =: (\widehat{Z}_{j1}, ..., \widehat{Z}_{jp})^\top,\ j = 1, ..., n \tag{4}$$

obtained by plugging estimators $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Omega}}_n$ of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ into (1). It is well known, however, that the matrix $\boldsymbol{\Omega}$ in (1) is not uniquely identified, and identification constraints need to be imposed which, without any loss of generality, actually define the parameter space for $\boldsymbol{\Omega}$. These constraints, which are closely related to the construction and statistical properties of $\widehat{\boldsymbol{\Omega}}_n$ are omitted here.

We now rapidly describe the proposed test statistics. Denote by

$$\varphi^{(n)}(\boldsymbol{t}) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\mathrm{i}\boldsymbol{t}^\top \widehat{\boldsymbol{Z}}_j}, \quad \boldsymbol{t} \in \mathbb{R}^p \tag{5}$$

and

$$\varphi_\ell^{(n)}(t_\ell) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\imath t_\ell \widehat{Z}_{j\ell}}, \quad t_\ell \in \mathbb{R}, \quad \ell = 1, \ldots, p \tag{6}$$

(where $\imath$ stands for the imaginary root of $-1$) the joint and marginal empirical CFs, respectively, of the estimated ICM residuals $\widehat{\boldsymbol{Z}}_j$, $j = 1, \ldots, n$.

Similarly, letting $\mathbf{R}_j^{(n)} \coloneqq (R_{j1}^{(n)}, \ldots, R_{jp}^{(n)})^\top$ where $R_{j\ell}^{(n)}$, $j = 1, \ldots, n$, stands for the rank of $\widehat{Z}_{j\ell}$ among $\widehat{Z}_{1\ell}, \ldots, \widehat{Z}_{n\ell}$, $\ell = 1, \ldots, p$, define the joint and marginal rank-based statistics

$$\underset{\sim}{\varphi}_{\mathbf{J}}^{(n)}(\boldsymbol{t}) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\imath \boldsymbol{t}^\top \mathbf{J}(\mathbf{R}_j^{(n)}/(n+1))}, \quad \boldsymbol{t} \in \mathbb{R}^p \tag{7}$$

and

$$\underset{\sim}{\varphi}_{\mathbf{J},\ell}^{(n)}(t_\ell) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\imath t_\ell J_\ell(R_{j\ell}^{(n)}/(n+1))}, \quad t_\ell \in \mathbb{R}, \quad \ell = 1, \ldots, p \tag{8}$$

where $\mathbf{J}(\mathbf{R}_j^{(n)}/(n+1)) \coloneqq (J_1(R_{j1}^{(n)}/(n+1)), \ldots, J_p(R_{jp}^{(n)}/(n+1)))^\top$ and $\mathbf{J} \coloneqq (J_1, \ldots, J_p)^\top$ denotes a $p$-tuple of *score functions* $J_\ell : (0,1) \to \mathbb{R}$. Clearly, (7) and (8) are the joint and marginal empirical CFs, respectively, of the *scored ranks* $J_\ell(R_{j\ell}^{(n)}/(n+1))$ of the estimated ICM residuals $\widehat{\boldsymbol{Z}}_j$, $j = 1, \ldots, n$.

Our tests are rejecting the null hypothesis $\mathcal{H}_0$ in (3) for large values of the test statistics

$$T_{n,W} \coloneqq n\int_{\mathbb{R}^p} |D_n(\boldsymbol{t})|^2 \, W(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \quad \text{and} \quad \underset{\sim}{T}_{n,\mathbf{J},W} \coloneqq n\int_{\mathbb{R}^p} |\underset{\sim}{D}_{n,\mathbf{J}}(\boldsymbol{t})|^2 \, W(\boldsymbol{t})\mathrm{d}\boldsymbol{t}, \tag{9}$$

(of the Cramér-von Mises type) where

$$D_n(\boldsymbol{t}) \coloneqq \varphi^{(n)}(\boldsymbol{t}) - \prod_{\ell=1}^{p} \varphi_\ell^{(n)}(t_\ell) \quad \text{and} \quad \underset{\sim}{D}_{n,\mathbf{J}}(\boldsymbol{t}) \coloneqq \underset{\sim}{\varphi}_{\mathbf{J}}^{(n)}(\boldsymbol{t}) - \prod_{\ell=1}^{p} \underset{\sim}{\varphi}_{\mathbf{J},\ell}^{(n)}(t_\ell) \tag{10}$$

with $\boldsymbol{t} \coloneqq (t_1, t_2, \ldots, t_p)^\top \in \mathbb{R}^p$ and some weight function $W : \boldsymbol{t} \mapsto W(\boldsymbol{t})$. Test statistics of the Kolmogorov-Smirnov type $S_n \coloneqq \sqrt{n}\sup_{\boldsymbol{t}\in\mathbb{R}^p} |D_n(\boldsymbol{t})|$ and $\underset{\sim}{S}_{n,\mathbf{J}} \coloneqq \sqrt{n}\sup_{\boldsymbol{t}\in\mathbb{R}^p} |\underset{\sim}{D}_{n,\mathbf{J}}(\boldsymbol{t})|$ could also be considered; see for instance Csörgő [1985]. However, computing a $\sup_{\boldsymbol{t}\in\mathbb{R}^p}$ runs into technical difficulties (already apparent in Csörgő [1985]) since the empirical CF converges weakly only over compact neighborhoods of $\mathbb{R}^p$ and this supremum has to be taken over some discrete grid. These drawbacks appear to have a direct impact on finite–sample powers, and Kolmogorov–Smirnov type tests typically are less powerful than their Cramér–von Mises counterparts (Henze et al. [2014]). We, therefore, only consider the latter.

# References


J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution-free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, 32:485–498, 1961.

S. Chakraborty and X. Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114:1638–1650, 2019.

A. Chen and P. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53:3625–3632, 2005.

F. Chen, S. G. Meintanis, and L. Zhu. On some characterizations and multidimensional criteria for testing homogeneity, symmetry and independence. *Journal of Multivariate Analysis*, 173:125–144, 2019.

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, 45:223–256, 2017.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Oxford, 2010.

S. Csörgő. Testing for independence by the empirical characteristic function. *Journal of Multivariate Analysis*, 16:290–299, 1985.

S. Csörgő and P. Hall. Estimable versions of Griffiths' measure of association. *Australian Journal of Statistics*, 24:296–308, 1982.

P. Deheuvels. An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11:102–113, 1981.

D. Edelmann, K. Fokianos, and M. Pitsillou. An updated literature review of distance correlation and its applications to time series. *International Statistical Review*, 87:237–262, 2019.

J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83:2195–2208, 2003.

Y. Fan, P. Lafaye de Micheaux, S. Penev, and D. Salopek. Multivariate nonparametric test of independence. *Journal of Multivariate Analysis*, 153:189–210, 2017.

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61:419–433, 1993.

A. Garcia-Ferrer, E. Gonzalez-Prieto, and D. Pena. A conditionally heteroskedastic independent factor model with an application to financial stock returns. *International Journal of Forecasting*, 28:70–93, 2012.

C. Genest, J. G. Nešlehová, B. Rémillard, and O. A. Murphy. Testing for independence in arbitrary distributions. *Biometrika*, 106:47–68, 2019.

P. Ghosal and B. Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *Annals of Statistics*, 50:1012–1037, 2022.

C. Gourieroux, A. Monfort, and J.-P. Renne. Statistical inference for independent component analysis: Application to structural VAR models. *Journal of Econometrics*, 196:111–126, 2017.

T. Hai. Estimation of volatility causality in structural autoregressions with heteroskedasticity using independent component analysis. *Statistical Papers*, 61:1–16, 2020.

J. Hájek and Z. Šidák. *Theory of Rank Tests*. Academic Press, New York, 1967.

M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *Annals of Statistics*, 49:1139–1165, 2021.

N. Henze, Z. Hlávka, and S. Meintanis. Testing for spherical symmetry via the empirical characteristic function. *Statistics*, 48:1282–1296, 2014.

H. Herwartz and S. Maxand. Nonparametric tests for independence: a review and comparative simulation study with an application to malnutrition in India. *Statistical Papers*, 61: 2175–2201, 2020.

A. Kankainen. *Consistent testing of total independence based on the empirical characteristic function*. PhD thesis, University of Jyväskylä, 1995.

A. Kankainen and N. Ushakov. A consistent modification of a test for independence based on the empirical characteristic function. *Journal of Mathematical Sciences*, 89:1486–1494, 1998.

D. S. Matteson and R. S. Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112:623–637, 2017.

S. G. Meintanis and G. Iliopoulos. Fourier methods for testing multivariate independence. *Computational Statistics & Data Analysis*, 52:1884–1895, 2008.

J. Miettinen, M. Matilainen, K. Nordhausen, and S. Taskinen. Extracting conditionally heteroskedastic components using independent component analysis. *Journal of Time Series Analysis*, 41:293–311, 2020.

H. Oja, D. Paindaveine, and S. Taskinen. Affine-invariant rank tests for multivariate independence in independent component models. *Electronic Journal of Statistics*, 10:2372–2419, 2016.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80:5–31, 2017.

A. Roy, S. Sarkar, A. K. Ghosh, and A. Goswami. On some consistent tests of mutual independence among several random vectors of arbitrary dimensions. *Statistics and Computing*, 30:1707–1723, 2020.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS–based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

H. Shi, M. Drton, and F. Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 117:395–410, 2022a.

H. Shi, M. Hallin, M. Drton, and F. Han. On universally consistent and fully distribution-free rank tests of vector independence. *Annals of Statistics*, 50:1933–1959, 2022b.

H. Shi, M. Drton, M. Hallin, and F. Han. Distribution-free tests of multivariate independence based on center-outward quadrant, Spearman, and Kendall statistics. *Bernoulli,* to appear, 2023.

G. J. Székely and M. L. Rizzo. *The Energy of Data and Distance Correlation.* CRC Press, Florida, USA, 2023.

S. S. Wilks. On the independence of $k$ sets of normally distributed statistical variables. *Econometrica*, 3:309–326, 1935.

# Séries temporelles 1

# Detecting the change points in a nonlinear time series models for weakly dependent observations

Echarif Elharfaoui [1] & Mohamed Salah Eddine Arrouch [2] & Joseph Ngatchou-Wandji [3]

[1,2] *Université Chouaib Doukkali, Faculté des Sciences, 24000 El Jadida, Maroc*
*elharfaoui.e@ucd.ac.ma, arrouch.m@ucd.ac.ma*
[3] *Institut Élie Cartan de Lorraine, Université de Rennes (EHESP), CEDEX, 54506 Vandoeuvre-Lès-Nancy, France, joseph.ngatchou-wandji@univ-lorraine.fr*

**Résumé.** Cet article étudie la détection de rupture d'une classe de modèles paramétriques conditionnels hétéroscédastiques non linéaires autorégressifs (CHARN). Les estimateurs conditionnels des moindres carrés (CLS) des paramètres sont définis et se révèlent cohérents. Un estimateur de la détection de rupture est défini. Sa consistance et sa distribution limite sont étudiées en détail.

**Mots-clés.** ruptures, modèles CHARN, moindres carrés conditionnels, mélange, tests

**Abstract.** This paper studies change-point detection of a class of parametric conditional heteroscedastic autoregressive nonlinear (CHARN) models. The conditional least-squares (CLS) estimators of the parameters are defined and are proved to be consistent. An estimator of the change-point location is defined. Its consistency and its limiting distribution are studied in detail.

**Keywords.** change-points, CHARN models, conditional least-squares, mixing, tests

## 1 Introduction

Detecting jumps in a series of real numbers and determining their number and locations is known in statistics as a change-point problem. This is usually solved by testing for the stationarity of the series and estimating the change locations when the null hypothesis of stationarity is rejected.

The study of conditional variations in financial and economic data receives particular attention as a result of its interest in hedging strategies and risk management.

The literature on change-points is vast. A popular alternative to using the likelihood ratio test was employed in Hinkley (1970,1972). Yao and Davis (1986) reviewed the asymptotic behavior of likelihood ratio statistics for testing a change in the mean in a series of iid Gaussian random variables. Csörgő and Horváth (1987) came up with statistics based on linear rank statistical processes with quantum scores. Chen and Gupta (1999) looked at detection tests and change-point estimation methods for models based on the normal distribution. The contribution of Horváth and Steinebach (2000) is related to the change in mean and variance. Antoch and Hušková (2001) proposed permutation tests for the location and scale

parameters of a law. Hušková and Meintanis (2006) developed a change-point test using the empirical characteristic functions. Zou et al. (2007) proposed several CUSUM approaches. Gombay (2008) proposed tests for change detection in the mean, variance, and autoregressive parameters of a p-order autoregressive model.

There are several types of changes depending on the temporal behavior of the series studied. The usual ones are abrupt change, gradual change, and intermittent change. In this paper, we focus on abrupt change in the conditional variance of *off-line* data issue from a class of CHARN models (see Härdle and Tsybakov (1997) and Härdle et al. (1998)). These models are of the most famous and significant ones in finance, which include many financial time series models. We suggest a hybrid estimation procedure, which combines CLS and non-parametric methods to estimate the change location. Indeed, conditional least-squares estimators own a computational advantage and require no knowledge of the innovation process.

The rest of the paper is organized as follows. Section 2 presents the class of models studied, the notation, the main assumptions and the main result on the CLS estimators of the parameters. Section 3 presents the change-point test and the change location LS estimation. The asymptotic distribution of the test statistic under the null hypothesis is investigated. The consistency rates are obtained for the change location estimator and its limit distribution is derived.

# 2 Model and Assumptions

We place ourselves in the framework where the observations at hand are assumed to be issued from the following CHARN $(p, p)$ model :

$$X_t = m\left(\rho; Z_{t-1}\right) + \sigma\left(\theta; Z_{t-1}\right)\varepsilon_t,\ t \in \mathbb{Z}, \tag{1}$$

where $p \in \mathbb{N}^* \bigcup \{\infty\}$ ; $m(\cdot)$ and $\sigma(\cdot)$ are two real-valued functions of known forms depending on unknown parameters $\rho$ and $\theta$, respectively ; for all $t \in \mathbb{Z}$, $Z_{t-1} = (X_{t-1}, X_{t-2}, \ldots, X_{t-p})^\intercal$ ; $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a sequence of stationary random variables with $\mathbb{E}\left(\varepsilon_t \mid Z_{t-1}\right) = 0$ and $\mathbb{V}\left(\varepsilon_t \mid Z_{t-1}\right) = 1$ such that $\varepsilon_t$ is independent of the $\sigma$-algebra $\mathcal{F}_{t-1} = \sigma\left(Z_k,\ k < t\right)$. The case $p = \infty$ is treated in Bardet and Kengne (2014) where the stationarity and the ergodicity of the process $(X_t)_{t\in\mathbb{Z}}$ is studied. Although we restrict to $p < \infty$, all the results stated here also hold for $p = \infty$.

Let $\psi = (\rho^\intercal, \theta^\intercal)^\intercal \in \Psi = \text{int}(\Theta) \times \text{int}(\widetilde{\Theta}) \subset \mathbb{R}^r \times \mathbb{R}^l$, the vector of the parameters of the model (1) and $\psi_0 = (\rho_0^\intercal, \theta_0^\intercal)^\intercal$ the true parameter vector. Denote by $\|M\|$ an appropriate norm of a vector or a matrix $M$. We assume that all the random variables in the whole text are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
We make the following assumptions :

($A_1$) The common fourth order moment of the $\varepsilon_t$ is finite.

($A_2$)

- The function $m(\cdot)$ is twice continuously differentiable, a.e., with respect to $\rho$ in some neighborhood $\mathcal{B}_1$ of $\rho_0$.
- The function $\sigma(.)$ is twice continuously differentiable, a.e., with respect to $\theta$ in some neighborhood $\mathcal{B}_2$ of $\theta_0$.

- There exists a positive function $\omega$ such that $\mathbb{E}\left(\omega^4\left(Z_0\right)\right) < \infty$, and

$$\max\left\{\sup_{\rho\in\mathrm{int}(\Theta)}\left|m(\rho;z)\right|,\ \sup_{\rho\in\mathrm{int}(\Theta)}\left\|\partial_\rho m(\rho;z)\right\|,\ \sup_{\rho\in\mathrm{int}(\Theta)}\left\|\partial^2_{\rho^2} m(\rho;z)\right\|\right\} \leqslant \omega(z)$$

$$\max\left\{\sup_{\theta\in\mathrm{int}(\widetilde{\Theta})}\left|\sigma(\theta;z)\right|,\ \sup_{\theta\in\mathrm{int}(\widetilde{\Theta})}\left\|\partial_\theta \sigma(\theta;z)\right\|,\ \sup_{\theta\in\mathrm{int}(\widetilde{\Theta})}\left\|\partial^2_{\theta^2}\sigma(\theta;z)\right\|\right\} \leqslant \omega(z).$$

($\mathrm{A}_3$) There exists a positive function $\beta$ such that $\mathbb{E}\left(\beta^4\left(Z_0\right)\right) < \infty$, and for all $\rho_1,\ \rho_2 \in \mathrm{int}(\Theta)$, and $\theta_1,\ \theta_2 \in \mathrm{int}(\widetilde{\Theta})$,

$$\begin{aligned}\max\Big\{&\left|m\left(\rho_1;z\right)-m\left(\rho_2;z\right)\right|,\left\|\partial_\rho m\left(\rho_1;z\right)-\partial_\rho m\left(\rho_2;z\right)\right\|,\\ &\left\|\partial^2_{\rho^2} m\left(\rho_1;z\right)-\partial^2_{\rho^2} m\left(\rho_2;z\right)\right\|,\left|\sigma\left(\theta_1;z\right)-\sigma\left(\theta_2;z\right)\right|,\\ &\left\|\partial_\theta \sigma\left(\theta_1;z\right)-\partial_\theta \sigma\left(\theta_2;z\right)\right\|,\left\|\partial^2_{\theta^2}\sigma\left(\theta_1;z\right)-\partial^2_{\theta^2}\sigma\left(\theta_2;z\right)\right\|\Big\}\\ &\leqslant \beta(z)\min\left\{\left\|\rho_1-\rho_2\right\|_2,\left\|\theta_1-\theta_2\right\|_2\right\}.\end{aligned}$$

($\mathrm{A}_4$) The sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ is stationary and satisfies either of the following two conditions :
- $\alpha$-mixing with mixing coefficient satisfying $\sum_{n\geq 1}\left[\alpha(n)\right]^{\delta/(2+\delta)} < \infty$ and $\mathbb{E}|\varepsilon_0|^{2+\delta} < \infty$ for some $\delta > 0$ ;
- $\phi$-mixing with mixing coefficient satisfying $\sum_{n\geqslant 1}\left[\phi(n)\right]^{1/2} < \infty$ and $\mathbb{E}\left|\varepsilon_0\right|^{4+\delta} < \infty$ for some $\delta > 0$.

The conditional mean and the conditional variance of $X_t$ are given, respectively, by

$$\mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right) = m\left(\rho;Z_{t-1}\right) \quad \text{and} \quad \mathbb{V}\left(X_t \mid \mathcal{F}_{t-1}\right) = \sigma^2\left(\theta;Z_{t-1}\right).$$

From these, one has that for all $z \in \mathbb{R}^p$,

$$\mathbb{E}\left(X_1 \mid Z_0 = z\right) = m\left(\rho;z\right) \quad \text{and} \quad \mathbb{E}\left(\left(X_1 - m\left(\rho;Z_0\right)\right)^2 \mid Z_0 = z\right) = \sigma^2\left(\theta;z\right).$$

Therefore, for any bounded measurable functions $g(\cdot)$ and $k(\cdot)$, we have

$$\mathbb{E}\left\{\left[X_1 - m\left(\rho;Z_0\right)\right]g(Z_0)\right\} = 0 \quad \text{and} \quad \mathbb{E}\left(\left\{\left[X_1 - m\left(\rho;Z_0\right)\right]^2 - \sigma^2\left(\theta;Z_0\right)\right\}k(Z_0)\right) = 0.$$

Without a loss of generality, in the following we take, for all $z \in \mathbb{R}^p$, $g(z) = k(z) = 1$. Now, given $X_{-p+1},\ldots,X_{-1},X_0,X_1,\ldots,X_n$ with $n \gg p$, we let $\mathbb{X}_n = (X_{-p+1},\ldots,X_{-1},X_0,X_1,\ldots,X_n)$ and consider the sequences of random functions

$$\begin{aligned} Q_n(\rho) = Q_n(\rho;\mathbb{X}_n) &= \sum_{t=1}^{n}\left(X_t - \mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right)\right)^2 = \sum_{t=1}^{n}\left(X_t - m\left(\rho;Z_{t-1}\right)\right)^2\\ S_n(\rho,\theta) = S_n(\rho,\theta;\mathbb{X}_n) &= \sum_{t=1}^{n}\left(\left(X_t - m\left(\rho;Z_{t-1}\right)\right)^2 - \sigma^2\left(\theta;Z_{t-1}\right)\right)^2.\end{aligned}$$

We have the following theorem :

**Theorem 1** *Under assumptions* $(A_1)$-$(A_3)$, *there exists a sequence of estimators* $\widehat{\psi}_n = \left(\widehat{\rho}_n^{\intercal}, \widehat{\theta}_n^{\intercal}\right)^{\intercal}$ *such that* $\widehat{\psi}_n \longrightarrow \psi_0$ *almost surely, and for any* $\epsilon > 0$, *there exists an event* $E$ *with* $\mathbb{P}(E) > 1-\epsilon$, *and a non-negative integer* $n_0$ *such that on* $E$, *for* $n > n_0$,

- $\dfrac{\partial Q_n}{\partial \rho}(\widehat{\rho}_n; \mathbb{X}_n) = 0$ *and* $Q_n(\rho; \mathbb{X}_n)$ *attains a relative minimum at* $\rho = \widehat{\rho}_n$ ;
- *assuming* $\widehat{\rho}_n$ *fixed,* $\dfrac{\partial S_n}{\partial \theta}\big(\widehat{\psi}_n; \mathbb{X}_n\big) = 0$ *and* $S_n\left((\widehat{\rho}_n, \theta); \mathbb{X}_n\right)$ *attains a relative minimum at* $\theta = \widehat{\theta}_n$.

**Sketch of the proof.** This result is an extension of Ngatchou-Wandji (2008) to the case $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a mixing martingale difference. □

# 3 Change-Point Study

## 3.1 Change-Point Test and Change Location Estimation

We essentially use the techniques of Bai (1994), who studied the estimation of the shift in the mean of a linear process by a LS method. We first consider the model (1) for known $\rho$, and $\sigma(\theta; Z_{t-1}) = \underline{\theta}\delta_0(Z_{t-1})$, for some known positive real-valued function $\delta_0(\cdot)$ defined on $\mathbb{R}^p$ and for an unknown positive real number $\underline{\theta}$. We wish to test

$$H_0 : \underline{\theta} = \vartheta_1 = \vartheta_2 \text{ over } t \leqslant n$$

against

$$H_1 : \underline{\theta} = \begin{cases} \vartheta_1, & t = 1, \ldots, t^* \\ \vartheta_2, & t = t^* + 1, \ldots, n. \end{cases} \qquad (\vartheta_1 \neq \vartheta_2)$$

where $\vartheta_1$, $\vartheta_2$ and $t^*$ are unknown parameters.
If we put

$$\sigma^2(\theta; Z_{t-1}) = \delta(Z_{t-1}) = \underline{\theta}\delta_0(Z_{t-1}) = \begin{cases} \vartheta_1\delta_0(Z_{t-1}), & t = 1, \ldots, t^* \\ \vartheta_2\delta_0(Z_{t-1}), & t = t^* + 1, \ldots, n. \end{cases} \qquad (\vartheta_1 \neq \vartheta_2)$$

and if we are also interested in estimating $\vartheta_1$, $\vartheta_2$ and the change location $t^*$, when $H_0$ is rejected. It is assumed that $t^* = [n\tau]$ for some $\tau \in (0,1)$, with $[x]$ standing for the integer part of any real number $x$. From (1), one can easily check that

$$\left(X_t - m(\rho; Z_{t-1})\right)^2 = \delta^2(Z_{t-1}) + \delta^2(Z_{t-1})\left(\varepsilon_t^2 - 1\right),\ t \in \mathbb{Z} \qquad (2)$$

and

$$\frac{\left(X_t - m(\rho; Z_{t-1})\right)^2}{\delta_0^2(Z_{t-1})} = \begin{cases} \vartheta_1^2\varepsilon_t^2, & t = 1, \ldots, t^* \\ \vartheta_2^2\varepsilon_t^2, & t = t^* + 1, \ldots, n. \end{cases} \qquad (\vartheta_1 \neq \vartheta_2)$$

from which we define the LS estimator $\widehat{t^*}$ of $t^*$ as follows :

$$\widehat{t^*} := \underset{1\leqslant k<n}{\arg\min} \left[\min_{\vartheta_1,\vartheta_2}\Big\{ \sum_{t=1}^{k}\left(W_t^2 - \vartheta_1^2\right)^2 + \sum_{t=k+1}^{n}\left(W_t^2 - \vartheta_2^2\right)^2 \Big\}\right], \qquad (3)$$

where $W_t = (X_t - m(\rho; Z_{t-1}))/\delta_0(Z_{t-1})$. Thus, the change location is estimated by minimizing the sum of squares of residuals among all possible sample splits. Letting

$$\overline{W}_k = \frac{1}{k}\sum_{t=1}^{k} W_t^2,\ \overline{W}_{n-k} = \frac{1}{n-k}\sum_{t=k+1}^{n} W_t^2 \text{ and } \overline{W} = \frac{1}{n}\sum_{t=1}^{n} W_t^2,$$

it is easily seen that for some $k$, the LS estimator of $\vartheta_1^2\,(t \leqslant k)$ and $\vartheta_2^2\,(t > k)$ are $\overline{W}_k$ and $\overline{W}_{n-k}$, respectively, and that (3) can be written as

$$\widehat{t}^* = \underset{1\leqslant k<n}{\arg\min} \left\{ \sum_{t=1}^{k} \left(W_t^2 - \overline{W}_k\right)^2 + \sum_{t=k+1}^{n} \left(W_t^2 - \overline{W}_{n-k}\right)^2 \right\} = \underset{1\leqslant k<n}{\arg\min}\ S_k^2. \tag{4}$$

Let $S^2 = \sum_{t=1}^{n} \left(W_t^2 - \overline{W}\right)^2$. A simple algebra gives

$$S^2 = S_k^2 + U_k, \tag{5}$$

where

$$U_k = k\left(\overline{W}_k - \overline{W}\right)^2 + (n-k)\left(\overline{W}_{n-k} - \overline{W}\right)^2. \tag{6}$$

From (4) and (5), we have

$$\widehat{t}^* = \underset{1\leqslant k<n}{\arg\min}\ \left(S^2 - U_k\right) = \underset{1\leqslant k<n}{\arg\max}\ U_k. \tag{7}$$

From (6), a simple algebraic computation gives the following alternative expression for $U_k$ :

$$U_k = \frac{n}{k(n-k)}\left(\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right)\right)^2 = \left(\sqrt{\frac{n}{k(n-k)}}\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right)\right)^2 = T_k^2. \tag{8}$$

It results from (7) and (8) that

$$\widehat{t}^* = \underset{1\leqslant k<n}{\arg\max}\ T_k^2 = \underset{1\leqslant k<n}{\arg\max}\ |T_k|. \tag{9}$$

Writing $T_k^2 = n\Delta_k^2$, it is immediate that

$$\Delta_k^2 = \frac{1}{k(n-k)}\left(\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right)\right)^2 = \frac{k}{(n-k)}\left(\overline{W}_k - \overline{W}\right)^2.$$

Simple computations give

$$\Delta_k^2 = \frac{k(n-k)}{n^2}\left(\overline{W}_{n-k} - \overline{W}_k\right)^2,$$

from which we have

$$\widehat{t}^* = \underset{1\leqslant k<n}{\arg\max}\ \Delta_k^2 = \underset{1\leqslant k<n}{\arg\max}\ |\Delta_k|. \tag{10}$$

The test statistic we use for testing $H_0$ against $H_1$ is a scale version of $\max\limits_{1\leqslant k\leqslant n-1} |T_k|$.
One can observe that under some conditions (e.g., $\varepsilon_t$ i.i.d. with $\varepsilon_t \sim \mathcal{N}(0,1)$), this statistic is the equivalent likelihood based test statistic for testing $H_0$ against $H_1$ (see, e.g., Hawkins (1977)). Let

$$C_k = \sum_{t=1}^{k} W_t^2, \quad C_{n-k} = \sum_{t=k+1}^{n} W_t^2 \quad \text{and} \quad C_n = \sum_{t=1}^{n} W_t^2. \tag{11}$$

By simple calculations, we obtain

$$T_k = \sqrt{\frac{n}{k(n-k)}} \sum_{t=1}^{k} \left(W_t^2 - \overline{W}\right) = \left(q\left(\frac{k}{n}\right)\right)^{-1} \left(\frac{1}{\sqrt{n}}\left(C_k - \frac{k}{n} C_n\right)\right), \tag{12}$$

where $q(\cdot)$ is a positive weight function defined for any $x \in (0,1)$ by $q(x) = \sqrt{x(1-x)}$.

## 3.2 Asymptotic Distribution of the Test Statistic

The study of the asymptotic distribution of the test statistic under $H_0$, is based on that of the process $\xi_n(\cdot)$ defined for any $s \in [0,1]$ by

$$\xi_n(s) = C_n(s) - sC_n(1), \tag{13}$$

where

$$C_n(s) = \begin{cases} 0 & \text{if } 0 \leqslant s < \dfrac{1}{n} \text{ and } 1 - \dfrac{1}{n} < s < 1 \\ \displaystyle\sum_{t=1}^{[ns]} W_t^2 & \text{if } \dfrac{1}{n} \leqslant s \leqslant 1 - \dfrac{1}{n} \\ \displaystyle\sum_{t=1}^{n} W_t^2 & \text{if } s = 1, \end{cases} \tag{14}$$

where we recall that $[ns]$ is the integer part of $ns$. For some $\delta \in (1/n, 1/2)$ and for any $s$ in $[\delta, 1-\delta]$, we define

$$T_n(s) = \frac{\xi_n(s)}{\sqrt{n} q(s)} \quad \text{and} \quad \Lambda_n = \max_{\delta \leqslant s \leqslant 1-\delta} \frac{|T_n(s)|}{\widehat{\sigma}_w}, \tag{15}$$

where $q(s) = \sqrt{s(1-s)}$ and $\widehat{\sigma}_w$ is any consistent estimator of

$$\sigma_w^2 = \mathbb{E}\left(W_1^2 - \mathbb{E}\left(W_1^2\right)\right)^2 + 2\sum_{t \geqslant 2} \mathbb{E}\left(\left(W_1^2 - \mathbb{E}\left(W_1^2\right)\right)\left(W_t^2 - \mathbb{E}\left(W_t^2\right)\right)\right).$$

For $\delta \in (0, 1/2)$, we denote by $D_\delta \equiv D([\delta, 1-\delta])$ the space of all right continuous functions with left limits on $[\delta, 1-\delta]$ endowed with the Skorohod metric. It is clear that $C_n(\cdot), \xi_n(\cdot) \in D_0$ and $T_n(\cdot) \in D_\delta$.

**Theorem 2** *Assume that the assumptions* $(A_1)$-$(A_4)$ *hold. Then, under* $H_0$, *we have*

$$\frac{\xi_n(s)}{\sigma_w \sqrt{n}} \xrightarrow{d} \widetilde{B}(s) \text{ in } D_0 \text{ as } n \longrightarrow \infty \quad \text{and} \quad \Lambda_n \xrightarrow{\mathcal{D}} \sup_{\delta \leqslant s \leqslant 1-\delta} \frac{|\widetilde{B}(s)|}{q(s)} \text{ as } n \longrightarrow \infty,$$

*where* $\left\{\widetilde{B}(s),\ 0 \leqslant s \leqslant 1\right\}$ *is a Brownian Bridge on* $[0,1]$.

It is worth noting that if the change occurs at the very beginning or at the very end of the data, we may not have sufficient observations to obtain consistent LSE estimators of the

parameters or these may not be unique. This is why we stress on the truncated version of the test statistic given in Zou et al. (2007) that we recall :

$$\Lambda_n = \max_{\frac{\nu}{n}\leqslant s\leqslant 1-\frac{\nu}{n}} \frac{|T_n(s)|}{\widehat{\sigma}_w}, \quad \text{for any } \ 1 \leqslant \nu < \frac{n}{2}.$$

By Theorem 2, it is easy to see that for any $1 \leqslant \nu < n/2$,

$$\sup_{\frac{\nu}{n}\leqslant s\leqslant 1-\frac{\nu}{n}} \frac{|T_n(s)|}{\widehat{\sigma}_w} - \sup_{\frac{\nu}{n}\leqslant s\leqslant 1-\frac{\nu}{n}} \frac{|\widetilde{B}(s)|}{q(s)} \xrightarrow{\mathcal{D}} 0 \text{ as } n \longrightarrow \infty,$$

which yields the asymptotic null distribution of the test statistic. With this, at level of significance $\alpha \in (0,1)$, $H_0$ is rejected if $\Lambda_n > C_{\alpha,n}$, where $C_{\alpha,n}$ is the $(1-\alpha)$-quantile of the distribution of the above limit. This quantile can be computed by observing that under $H_0$, for larger values of $n$, one has

$$\alpha = \mathbb{P}\bigg(\sup_{\frac{\nu}{n}\leqslant s\leqslant 1-\frac{\nu}{n}} \frac{|T_n(s)|}{\widehat{\sigma}_w} > C_{\alpha,n}\bigg) \approx \mathbb{P}\bigg(\sup_{h_\nu(n)\leqslant s\leqslant 1-h_\nu(n)} \frac{|\widetilde{B}(s)|}{q(s)} > C_{\alpha,n}\bigg), \quad \text{where } \ h_\nu(n) = \frac{\nu}{n}.$$

From the following relation (1.3.26) of Csörgő and Horváth (1997), for each $h_\nu(n) > 0$, and for larger real number $x$, we have

$$\begin{aligned} \mathbb{P}\bigg\{\sup_{h_\nu(n)\leqslant s\leqslant 1-h_\nu(n)} \frac{|\widetilde{B}(s)|}{q(s)} \geqslant x\bigg\} \ &= \ \frac{1}{\sqrt{2\pi}} x \exp\left(\frac{-x^2}{2}\right)\left[\ln\left(\frac{(1-h_\nu(n))^2}{h_\nu^2(n)}\right)\right. \\ &\quad \left. -\frac{1}{x^2}\ln\left(\frac{(1-h_\nu(n))^2}{h_\nu^2(n)}\right) + \frac{4}{x^2} + \mathcal{O}\Big(\frac{1}{x^4}\Big)\right], \end{aligned} \tag{16}$$

which gives an approximation of the tail distribution of $\sup_{h_\nu(n)\leqslant s\leqslant 1-h_\nu(n)} |\widetilde{B}(s)|/q(s)$. Thus, using $\widehat{\sigma}_w$, an estimation of $C_{\alpha,n}$ can be obtained from this approximation. Monte Carlo simulations are often carried out to obtain accurate approximations of $C_{\alpha,n}$. In this purpose, it is necessary to make a good choice of $\nu$. We selected $\nu = 0.9 \times n^{4/5}$ as our option, which we found to be a suitable choice for all the cases we examined. But, to avoid the difficulties associated with the computation of $C_{\alpha,n}$, a decision can also be taken by using the $p$-value method as in Ngatchou-Wandji et al. (2022). That is using the approximation (16), reject $H_0$ if

$$\mathbb{P}\big(|\widetilde{B}(s)|q(s) > \Lambda_n\big) \leq \alpha.$$

This idea is used in the simulation.

## 3.3 Rate of Convergence of the Change Location Estimator

For the study of the estimator $\widehat{t^*}$, we let $\kappa = \kappa_n = \vartheta_2^2 - \vartheta_1^2$ and assume without loss of generality that $\kappa_n > 0$ $(\vartheta_2 > \vartheta_1)$, $\kappa_n \longrightarrow 0$ as $n \longrightarrow \infty$ and that the unknown change point $t^*$ depends on the sample size $n$. We have the following result :

**Theorem 3** *Assume that* $(A_4)$ *is satisfied,* $t^*/n \in (a, 1-a)$ *for some* $0 < a < 1/2$, $t^* = [n\tau]$ *for some* $\tau \in (0,1)$ *and as* $n \longrightarrow \infty$, $\kappa_n \longrightarrow 0$ *and* $\dfrac{\kappa_n\sqrt{n}}{\sqrt{\ln n}} \longrightarrow \infty$. *Then, we have*

$$\widehat{t^*} - t^* = O_{\mathbb{P}}\Big(\frac{1}{\kappa_n^2}\Big),$$

## 3.4 Limit Distribution of the Location Estimator

In this section, we study the asymptotic behavior of the location estimator. We make the additional assumptions that $\kappa_n >> n^{-\frac{1}{2}}$ and that as $n \longrightarrow \infty$,

$$\frac{\kappa_n\sqrt{n}}{\sqrt{\ln n}} \longrightarrow \infty \text{ and } n^{\frac{1}{2}-\zeta}\kappa_n \longrightarrow \infty \text{ for some } \zeta \in \left(0, \frac{1}{2}\right).$$

By (10), we have

$$\widehat{t^*} \;=\; \underset{1\leqslant k<n}{\arg\max}\, n\left(\Delta_k^2 - \Delta_{t^*}^2\right). \tag{17}$$

To derive the limiting distribution of $\widehat{t^*}$, we study the behavior of $n\left(\Delta_k^2 - \Delta_{t^*}^2\right)$ for those $k$s in the neighborhood of $t^*$ such that $k = [t^* + r\kappa_n^{-2}]$, where $r$ varies in an arbitrary bounded interval $[-N, N]$. For this purpose, we define

$$P_n(r) := n\left\{\Delta_n^2\left(\left[t^* + r\kappa_n^{-2}\right]\right) - \Delta_n^2(t^*)\right\},$$

where $\Delta_n(r) = \Delta_{[r]}$. In addition, we define the two-sided standard Wiener process $\{B^*(r),\ r \in \mathbb{R}\}$ as follows :

$$B^*(r) := \begin{cases} B_1(-r) & \text{if } r < 0 \\ B_2(r) & \text{if } r \geqslant 0, \end{cases}$$

where $B_i(r),\ i = 1,\ 2$ are two independent standard Wiener processes defined on $[0, \infty)$ with $B_i(0) = 0,\ i = 1,\ 2$.

First, we identify the limit of the process $P_n(r)$ on $|r| \leqslant N$ for every given $N > 0$. We denote by $C([-N, N])$ the space of all continuous functions on $[-N, N]$ endowed with the uniform metric.

**Proposition 1** *Assume that $(A_4)$ holds, that $t^* = [n\tau]$ for some $\tau \in (0,1)$ and that as $n \longrightarrow \infty$, $\kappa_n \longrightarrow 0$ and $\frac{\kappa_n\sqrt{n}}{\sqrt{\ln n}} \longrightarrow \infty$. Then, for every $0 < N < \infty$, the process $P_n(r)$ converges weakly in $C([-N, N])$ to the process $P(r) = 2\left\{\sigma_w B^*(r) - \frac{1}{2}|r|\right\}$, where $B^*(\cdot)$ is the two-sided standard Wiener process defined above.*

The above results make it possible to achieve a weak convergence result for $n\left(\Delta_k^2 - \Delta_{t^*}^2\right)$ and then apply the Argmax-Continuous Mapping Theorem (Argmax-CMT). We have :

**Theorem 4** *Assume that $(A_4)$ is satisfied, that $t^* = [n\tau]$ for some $\tau \in (0,1)$ and as $n \longrightarrow \infty$, $\kappa_n \longrightarrow 0$ and $\frac{\kappa_n\sqrt{n}}{\sqrt{\ln n}} \longrightarrow \infty$. Then we have $\frac{\kappa_n^2\left(\widehat{t^*} - t^*\right)}{\sigma_w^2} \xrightarrow{\mathcal{D}} \mathcal{S}$,*
*where $\mathcal{S} := \arg\max\left\{B^*(u) - \frac{1}{2}|u|,\ u \in \mathbb{R}\right\}$.*

This result yields the asymptotic distribution of the change location estimator. Bhattacharya (1987), Picard (1985) and Yao (1987) investigated the density function of the random variable $\mathcal{S}$ (see Lemma 1.6.3 of Csörgő and Horváth (1987) for more details). They also showed that $\mathcal{S}$ has a symmetric (with respect to 0) probability density function $\gamma(\cdot)$ defined for any $x \in \mathbb{R}$ by $\gamma(x) = \frac{3}{2}\exp(|x|)\Phi\left(\frac{-3}{2}\sqrt{|x|}\right) - \frac{1}{2}\Phi\left(\frac{-1}{2}\sqrt{|x|}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal variable. From this result, a confidence interval for the change-point location can be obtained, if one has consistent estimates of $\kappa_n^2$ and $\sigma_w^2$. With $\widehat{t^*}$, consistent estimates of $\vartheta_1^2$ and $\vartheta_2^2$ are given, respectively, by

$$\widehat{\vartheta}_1^2 = \overline{W}_{\widehat{t}^*} = \frac{1}{\widehat{t}^*}\sum_{t=1}^{\widehat{t}^*} W_t^2 \text{ and } \widehat{\vartheta}_2^2 = \overline{W}_{n-\widehat{t}^*} = \frac{1}{n-\widehat{t}^*}\sum_{t=\widehat{t}^*+1}^{n} W_t^2.$$

Thus, a consistent estimate of $\kappa_n^2$ is given by

$$\widehat{\kappa}_n^2 = \frac{1}{n-\widehat{t}^*}\sum_{t=\widehat{t}^*+1}^{n} W_t^2 - \frac{1}{\widehat{t}^*}\sum_{t=1}^{\widehat{t}^*} W_t^2.$$

A consistent estimator of $\sigma_w^2$ that we denote by $\widehat{\sigma}_w^2$ can be easily obtained by taking its empirical counterpart. So, at risk $\alpha \in (0,1)$, letting $q_{1-\frac{\alpha}{2}}$ be the quantile of order $1-\frac{\alpha}{2}$ of the distribution of the random variable $\mathcal{S}$, an asymptotic confidence interval for $t^*$ is given by

$$\text{CI} = \widehat{t}^* \pm \left( \left[ q_{1-\frac{\alpha}{2}} \frac{\widehat{\sigma}_w^2}{\widehat{\kappa}_n^2} \right] + 1 \right).$$

**Remark 1** *In the case that the parameter $\rho$ is unknown, it can be estimated by the CLS method (see Section 2), and be substituted for its estimator in $W_t$. Indeed, one can easily show that*

$$\frac{1}{k}\sum_{t=1}^{k} W_t^2 = \frac{1}{k}\sum_{t=1}^{k} \widehat{W}_t^2 + o_{\mathbb{P}}(1) \text{ and } \frac{1}{n-k}\sum_{t=k+1}^{n} W_t^2 = \frac{1}{n-k}\sum_{t=k+1}^{n} \widehat{W}_t^2 + o_{\mathbb{P}}(1),$$

*where for any $t = 1, \ldots, n$, $\widehat{W}_t = \left(X_t - m\left(\widehat{\rho}_n; Z_{t-1}\right)\right)/\delta_0\left(Z_{t-1}\right)$ and $\widehat{\rho}_n$ is the conditional least squares estimators of $\rho$ obtained from Theorem 1. Hence, the same techniques as in the case where $\rho$ is known can be used.*

# Bibliography


Antoch, J. and Hušková, M. (2001), Permutation tests in change point analysis, *Statistics & Probability Letters*, 53, pp. 37-46.

Arrouch, M. S.-E., Elharfaoui E. and Ngatchou-Wandji J. (2023), Change-Point Detection in the Volatility of Conditional Heteroscedastic Autoregressive Nonlinear Models, *Mathematics*, 11(18), `https://doi.org/10.3390/math11184018`

Bai, J. (1994), Least squares estimation of a shift in linear processes, *Journal of Time Series Analysis*, 15, pp. 453-472.

Bardet, J.M. and Kengne, W. (2014), Monitoring procedure for parameter change in causal time series, *Journal of Multivariate Analysis*, 125, pp. 204-221.

Bhattacharya, P.K. (1987), Maximum likelihood estimation of a change-point in the distribution of independent random variables : general multiparameter case, *Journal of Multivariate Analysis*, 23, pp. 183-208.

Chen, J. and Gupta, A. (1999), Change point analysis of a Gaussian model, *Statistical Papers*, 40, pp. 323-333.

Csörgő M. and Horváth L. (1997), *Limit Theorems in Change-Point Analysis*, Wiley, Chichester, UK.

Csörgő, M. and Horváth, L. (1987), Nonparametric tests for the changepoint problem, *Journal of Statistical Planning and Inference*, 17, pp. 1-9.

Gombay, E. (2008), Change detection in autoregressive time series, *Journal of Multivariate Analysis*, 99, pp. 451-464.

Hawkins, D.M. (1977), Testing a sequence of observations for a shift in location, *Journal of American Statistical Association*, 72, pp. 180-186.

Härdle, W. and Tsybakov, A. (1997), Local polynomial estimators of the volatility function in nonparametric autoregression, *Journal of Econometrics*, 81, pp. 223-242.

Härdle, W., Tsybakov, A. and Yang, L. (1998), Nonparametric vector autoregression, *Journal of Statistical Planning and Inference*, 68, pp. 221-245.

Hinkley, D.V. (1970), Inference about the change-point in a sequence of random variables, *Biometrika*, 57, pp. 1-17.

Hinkley, D.V. (1972), Time-ordered classification, *Biometrika*, 59, pp. 509-523.

Horváth, L. and Steinebach, J. (2000), Testing for changes in the mean or variance of a stochastic process under weak invariance, *Journal of Statistical Planning and Inference*, 91, pp. 365-376.

Hušková, M. and Meintanis, S.G. (2006), Change point analysis based on empirical characteristic functions, *Metrika*, 63, pp. 145-168.

Inclan, C. and Tiao, G.C. (1994), Use of cumulative sums of squares for retrospective detection of changes of variance, *Journal of American Statistical Association*, 89, pp. 913-923.

Lombard, F. (1987), Rank tests for changepoint problems, *Biometrika*, 74, pp. 615-624.

Ngatchou-Wandji, J. (2008), Estimation in a class of nonlinear heteroscedastic time series models, *Electronic Journal of Statistics*, 2, pp. 40-62.

Ngatchou-Wandji, J., Elharfaoui, E. and Harel, M. (2022), On change-points tests based on two-samples U-Statistics for weakly dependent observations, *Statistical Papers*, 63, pp. 287-316.

Picard, D. (1985), Testing and estimating change-points in time series, *Advances in Applied Probability*, 17, pp. 841-867.

Yao, Y.C. (1987), Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables, *Annals of Statistics*, 15, pp. 1321-1328.

Yao, Y.C. and Davis, R.A. (1986), The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates, *Sankhyā : The Indian Journal of Statistics, Series A*, pp. 339-353.

Zou, C, Liu, Y., Qin, P. and Wang, Z. (2007), Empirical likelihood ratio test for the change-point problem, *Statistics & Probability Letters*, 77, pp. 374-382.

# Detecting and estimating changepoints in nonlinear autoregressive models using simulated data

Echarif Elharfaoui [1] & Mohamed Salah Eddine Arrouch [2]

[1,2] *Université Chouaib Doukkali, Faculté des Sciences, 24000 El Jadida, Maroc*
*elharfaoui.e@ucd.ac.ma, arrouch.m@ucd.ac.ma*

**Résumé.** Cet article présente une étude par simulation de la détection de ruptures des modèles paramétriques conditionnels hétéroscédastiques non linéaires autorégressifs (CHARN). Une expérience de simulation est réalisée et appliquée à des ensembles de données réelles.

**Mots-clés.** modèles CHARN, moindres carrés conditionnels, ruptures, tests

**Abstract.** This paper presents a simulation study of change-point detection of parametric conditional heteroscedastic autoregressive nonlinear (CHARN) models. A simulation experiment is carried and applied to the sets of real data.

**Keywords.** CHARN models, conditional least-squares, change-points, tests

## 1 Introduction

Volatility models play a crucial role in the analysis of time series because they provide an overview of the degree of uncertainty and risk associated with a set of data. It refers to the phenomenon in which the variance of the terms of error in a time series model is not constant in time, which can complicate our statistical analyzes.

Detection of heteroscedasticity is the first crucial step to remedy it. Various methods to detect and discuss this volatility model. The study of conditional variations in financial and economic data receives particular attention as a result of its interest in hedging strategies and risk management. These models are of the most famous and significant ones in finance, which include many financial time series models. We suggest a hybrid estimation procedure, which combines CLS and non-parametric methods to estimate the change location. Indeed, conditional least-squares estimators own a computational advantage and require no knowledge of the innovation process.

The literature on change-points is vast. A popular alternative to using the likelihood ratio test was employed in Hinkley (1970,1972). Yao and Davis (1986) reviewed the asymptotic behavior of likelihood ratio statistics for testing a change in the mean in a series of iid Gaussian random variables. Csörgő and Horváth (1987) came up with statistics based on linear rank statistical processes with quantum scores. Chen and Gupta (1999) looked at detection and autoregressive parameters of a p-order autoregressive model.

There are several types of changes depending on the temporal behavior of the series studied. The usual ones are abrupt change, gradual change, and intermittent change. In this

paper, we focus on abrupt change in the conditional variance of *off-line* data issue from a class of CHARN models (see Härdle and Tsybakov (1997) and Härdle et al. (1998)). These models are of the most famous and significant ones in finance, which include many financial time series models.

The rest of the paper is organized as follows. Section 2 presents the definitions and concepts, the main assumptions on the CLS estimators of the parameters. Section 3 presents the simulation results from a few simple time series models. They are applied on real data sets.

# 2 Definitions and concepts

## 2.1 Model and Conditional Least-Squares Estimation

We place ourselves in the framework where the observations at hand are assumed to be issued from the following CHARN $(p, p)$ model :

$$X_t = m\left(\rho; Z_{t-1}\right) + \sigma\left(\theta; Z_{t-1}\right)\varepsilon_t,\ t \in \mathbb{Z}, \tag{1}$$

where $p \in \mathbb{N}^* \bigcup \{\infty\}$ ; $m(\cdot)$ and $\sigma(\cdot)$ are two real-valued functions of known forms depending on unknown parameters $\rho$ and $\theta$, respectively ; for all $t \in \mathbb{Z}$, $Z_{t-1} = (X_{t-1}, X_{t-2}, \ldots, X_{t-p})^\intercal$ ; $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a sequence of stationary random variables with $\mathbb{E}\left(\varepsilon_t \mid Z_{t-1}\right) = 0$ and $\mathbb{V}\left(\varepsilon_t \mid Z_{t-1}\right) = 1$ such that $\varepsilon_t$ is independent of the $\sigma$-algebra $\mathcal{F}_{t-1} = \sigma\left(Z_k,\ k < t\right)$. The case $p = \infty$ is treated in Bardet and Kengne (2014) where the stationarity and the ergodicity of the process $(X_t)_{t\in\mathbb{Z}}$ is studied. Although we restrict to $p < \infty$, all the results stated here also hold for $p = \infty$.

Let $\psi = (\rho^\intercal, \theta^\intercal)^\intercal \in \Psi = \text{int}(\Theta) \times \text{int}(\widetilde{\Theta}) \subset \mathbb{R}^r \times \mathbb{R}^l$, the vector of the parameters of the model (1) and $\psi_0 = (\rho_0^\intercal, \theta_0^\intercal)^\intercal$ the true parameter vector. Denote by $\|M\|$ an appropriate norm of a vector or a matrix $M$. We assume that all the random variables in the whole text are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, assuming, the common fourth order moment of the $\varepsilon_t$ is finite and regularity conditions for $m(\cdot)$ and $\sigma(\cdot)$.
The conditional mean and the conditional variance of $X_t$ are given, respectively, by

$$\mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right) = m\left(\rho; Z_{t-1}\right) \text{ and } \mathbb{V}\left(X_t \mid \mathcal{F}_{t-1}\right) = \sigma^2\left(\theta; Z_{t-1}\right).$$

From these, one has that for all $z \in \mathbb{R}^p$,

$$\mathbb{E}\left(X_1 \mid Z_0 = z\right) = m\left(\rho; z\right) \quad \text{and} \quad \mathbb{E}\left(\left(X_1 - m\left(\rho; Z_0\right)\right)^2 \mid Z_0 = z\right) = \sigma^2\left(\theta; z\right).$$

Therefore, for any bounded measurable functions $g(\cdot)$ and $k(\cdot)$, we have

$$\mathbb{E}\left\{\left[X_1 - m\left(\rho; Z_0\right)\right] g(Z_0)\right\} = 0 \quad \text{and} \quad \mathbb{E}\left(\left\{\left[X_1 - m\left(\rho; Z_0\right)\right]^2 - \sigma^2\left(\theta; Z_0\right)\right\} k(Z_0)\right) = 0.$$

Without a loss of generality, in the following we take, for all $z \in \mathbb{R}^p$, $g(z) = k(z) = 1$. Now, given $X_{-p+1}, \ldots, X_{-1}, X_0, X_1, \ldots, X_n$ with $n \gg p$, we let $\mathbb{X}_n = (X_{-p+1}, \ldots, X_{-1}, X_0, X_1, \ldots, X_n)$ and consider the sequences of random functions

$$\begin{aligned} Q_n(\rho) = Q_n(\rho; \mathbb{X}_n) &= \sum_{t=1}^{n} \left(X_t - \mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right)\right)^2 = \sum_{t=1}^{n} \left(X_t - m\left(\rho; Z_{t-1}\right)\right)^2 \\ S_n(\rho, \theta) = S_n(\rho, \theta; \mathbb{X}_n) &= \sum_{t=1}^{n} \left(\left(X_t - m\left(\rho; Z_{t-1}\right)\right)^2 - \sigma^2\left(\theta; Z_{t-1}\right)\right)^2. \end{aligned}$$

Under regularity conditions, we have

- $\dfrac{\partial Q_n}{\partial \rho}\left(\widehat{\rho}_n; \mathbb{X}_n\right) = 0$ and $Q_n(\rho; \mathbb{X}_n)$ attains a relative minimum at $\rho = \widehat{\rho}_n$;
- assuming $\widehat{\rho}_n$ fixed, $\dfrac{\partial S_n}{\partial \theta}\left(\widehat{\psi}_n; \mathbb{X}_n\right) = 0$ and $S_n\left(\left(\widehat{\rho}_n, \theta\right); \mathbb{X}_n\right)$ attains a relative minimum at $\theta = \widehat{\theta}_n$.

## 2.2 Change-Point Test and Change Location Estimation

We essentially use the techniques of Bai (1994), who studied the estimation of the shift in the mean of a linear process by a LS method. We first consider the model (1) for known $\rho$, and $\sigma\left(\theta; Z_{t-1}\right) = \underline{\theta}\delta_0\left(Z_{t-1}\right)$, for some known positive real-valued function $\delta_0(\cdot)$ defined on $\mathbb{R}^p$ and for an unknown positive real number $\underline{\theta}$. We wish to test $H_0 : \underline{\theta} = \vartheta_1 = \vartheta_2$ over $t \leqslant n$ against

$$H_1 : \underline{\theta} = \begin{cases} \vartheta_1, & t = 1, \ldots, t^* \\ \vartheta_2, & t = t^* + 1, \ldots, n. \end{cases} \qquad (\vartheta_1 \neq \vartheta_2)$$

where $\vartheta_1$, $\vartheta_2$ and $t^*$ are unknown parameters.

We are also interested in estimating $\vartheta_1$, $\vartheta_2$ and the change location $t^*$, when $H_0$ is rejected. It is assumed that $t^* = [n\tau]$ for some $\tau \in (0, 1)$, with $[x]$ standing for the integer part of any real number $x$. From (1), one can easily check that

$$\left(X_t - m\left(\rho; Z_{t-1}\right)\right)^2 = \delta^2\left(Z_{t-1}\right) + \delta^2\left(Z_{t-1}\right)\left(\varepsilon_t^2 - 1\right),\ t \in \mathbb{Z} \tag{2}$$

from which we define the LS estimator $\widehat{t^*}$ of $t^*$ as follows :

$$\widehat{t^*} := \underset{1 \leqslant k < n}{\arg\min} \left[ \min_{\vartheta_1, \vartheta_2} \left\{ \sum_{t=1}^{k} \left(W_t^2 - \vartheta_1^2\right)^2 + \sum_{t=k+1}^{n} \left(W_t^2 - \vartheta_2^2\right)^2 \right\} \right], \tag{3}$$

where $W_t = \left(X_t - m(\rho; Z_{t-1})\right) / \delta_0\left(Z_{t-1}\right)$. Thus, the change location is estimated by minimizing the sum of squares of residuals among all possible sample slits. Letting

$$\overline{W}_k = \frac{1}{k}\sum_{t=1}^{k} W_t^2,\ \overline{W}_{n-k} = \frac{1}{n-k}\sum_{t=k+1}^{n} W_t^2 \text{ and } \overline{W} = \frac{1}{n}\sum_{t=1}^{n} W_t^2,$$

it is easily seen that for some $k$, the LS estimator of $\vartheta_1^2\,(t \leqslant k)$ and $\vartheta_2^2\,(t > k)$ are $\overline{W}_k$ and $\overline{W}_{n-k}$, respectively, and that (3) can be written as

$$\widehat{t^*} = \underset{1 \leqslant k < n}{\arg\min} \left\{ \sum_{t=1}^{k} \left(W_t^2 - \overline{W}_k\right)^2 + \sum_{t=k+1}^{n} \left(W_t^2 - \overline{W}_{n-k}\right)^2 \right\} = \underset{1 \leqslant k < n}{\arg\min}\, S_k^2. \tag{4}$$

Let $S^2 = \sum_{t=1}^{n} \left(W_t^2 - \overline{W}\right)^2$. A simple algebra gives

$$S^2 = S_k^2 + U_k, \tag{5}$$

where

$$U_k = k\left(\overline{W}_k - \overline{W}\right)^2 + (n-k)\left(\overline{W}_{n-k} - \overline{W}\right)^2. \tag{6}$$

From (4) and (5), we have

$$\widehat{t^*} = \underset{1\leqslant k<n}{\arg\min}\ \left(S^2 - U_k\right) = \underset{1\leqslant k<n}{\arg\max}\ U_k. \tag{7}$$

From (6), a simple algebraic computation gives the following alternative expression for $U_k$ :

$$U_k = \frac{n}{k(n-k)}\left(\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right)\right)^2 = \left(\sqrt{\frac{n}{k(n-k)}}\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right)\right)^2 = T_k^2. \tag{8}$$

It results from (7) and (8) that

$$\widehat{t^*} = \underset{1\leqslant k<n}{\arg\max}\ T_k^2 = \underset{1\leqslant k<n}{\arg\max}\ |T_k|. \tag{9}$$

Writing $T_k^2 = n\Delta_k^2$, it is immediate that

$$\Delta_k^2 = \frac{1}{k(n-k)}\left(\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right)\right)^2 = \frac{k}{(n-k)}\left(\overline{W}_k - \overline{W}\right)^2.$$

Simple computations give $\Delta_k^2 = \dfrac{k(n-k)}{n^2}\left(\overline{W}_{n-k} - \overline{W}_k\right)^2$, from which we have

$$\widehat{t^*} = \underset{1\leqslant k<n}{\arg\max}\ \Delta_k^2 = \underset{1\leqslant k<n}{\arg\max}\ \left|\Delta_k\right|. \tag{10}$$

The test statistic we use for testing $H_0$ against $H_1$ is a scale version of $\max\limits_{1\leqslant k\leqslant n-1}|T_k|$.

One can observe that under some conditions (e.g., $\varepsilon_t$ i.i.d. with $\varepsilon_t \sim \mathcal{N}(0,1)$), this statistic is the equivalent likelihood based test statistic for testing $H_0$ against $H_1$ (see, e.g., Hawkins (1977)). Let

$$C_k = \sum_{t=1}^{k} W_t^2, \quad C_{n-k} = \sum_{t=k+1}^{n} W_t^2 \quad \text{and} \quad C_n = \sum_{t=1}^{n} W_t^2. \tag{11}$$

By simple calculations, we obtain

$$T_k = \sqrt{\frac{n}{k(n-k)}}\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right) = \left(q\left(\frac{k}{n}\right)\right)^{-1}\left(\frac{1}{\sqrt{n}}\left(C_k - \frac{k}{n}C_n\right)\right), \tag{12}$$

where $q(\cdot)$ is a positive weight function defined for any $x \in (0,1)$ by $q(x) = \sqrt{x(1-x)}$.

The study of the asymptotic distribution of the test statistic under $H_0$, is based on that of the the process $\xi_n(\cdot)$ defined for any $s \in [0,1]$ by $\xi_n(s) = C_n(s) - sC_n(1)$,

For $\delta \in (0, 1/2)$, we denote by $D_\delta \equiv D([\delta, 1-\delta])$ the space of all right continuous functions with left limits on $[\delta, 1-\delta]$ endowed with the Skorohod metric. It is clear that $C_n(\cdot), \xi_n(\cdot) \in D_0$ and $T_n(\cdot) \in D_\delta$.

For the study of the estimator $\widehat{t^*}$, we let $\kappa = \kappa_n = \vartheta_2^2 - \vartheta_1^2$ and assume without loss of generality that $\kappa_n > 0$ $(\vartheta_2 > \vartheta_1)$, $\kappa_n \longrightarrow 0$ as $n \longrightarrow \infty$ and that the unknown change point $t^*$ depends on the sample size $n$. We have the following result :

**Theorem 1** *Assume that mixing is satisfied, $t^*/n \in (a, 1-a)$ for some $0 < a < 1/2$, $t^* = [n\tau]$ for some $\tau \in (0,1)$ and as $n \longrightarrow \infty$, $\kappa_n \longrightarrow 0$ and $\dfrac{\kappa_n\sqrt{n}}{\sqrt{\ln n}} \longrightarrow \infty$. Then, we have*

$$\widehat{t^*} - t^* = O_{\mathbb{P}}\left(\frac{1}{\kappa_n^2}\right),$$

# 3 Practical Consideration

In this section we perform numerical simulations to evaluate the performances of our methods and these are applied to two sets of real data. We start with the presentation of the results of numerical simulations found with the software R, version 4.1.1. The trials are based on 1000 replications of observations of lengths $n = 500,\ 1000,\ 5000$ and 10000 generated from the model (1) for $\rho = (\rho_0, \rho_1, \rho_2)^\intercal$ ; $\theta = (\theta_0, \theta_1, \theta_2)^\intercal$ ; $m(\rho; x) = (\rho_0 + \rho_1 \exp(-\rho_2 x^2))\, x$; $\sigma(\theta; x) = \underline{\theta}\delta_0(x)$ with $\delta_0(x) = \sqrt{\theta_0^2 + \theta_1^2 x^2 \exp(-\theta_2 x^2)}$; $\rho_2 > 0,\ \rho_0\rho_1 \geqslant 0,\ \theta_2 \geqslant 0$ and $0 < \underline{\theta}^2\theta_1^2 < 1$; $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a white noise with density function $f$. We also assume the sufficient condition $|\rho_0| + |\rho_1| + |\underline{\theta}\theta_1| + 2\rho_0\rho_1 < 1$, to ensure the strict stationarity and ergodicity of the process $(X_t)_{t\in\mathbb{Z}}$ (see, e.g., Theorem 3.2.11 of Taniguchi and Kakizawa (2000), p. 86 and Ngatchou-Wandji (2005), p. 5). The noise densities $f$ that we employed were Gaussian.
The change-point location is estimated using the following algorithm :

**Algorithm 1** Change-point location estimation

1: **for** $i = 1, \ldots, 1000$ **do**
2: **for** $t = 1, \ldots, n$ **do** $W_t = (X_t - m(\rho; X_{t-1}))\,/\delta_0(X_{t-1})$
3: **end for**
4: $\overline{W} = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n} W_t^2$
5: **for** $k = 1, \ldots, n-1$ **do** $T_k = \sqrt{\dfrac{n}{k(n-k)}}\displaystyle\sum_{t=1}^{k}\left(W_t^2 - \overline{W}\right)$
6: **end for**
7: Compute $\widehat{t}_i^{\,*} = \underset{1\leqslant k<n}{\arg\max}\,|T_k|$ (that is the value of $k$ for which $|T_k|$ is the largest)
8: **end for**
9: Compute $L = \left(\widehat{t}_1^{\,*} + \widehat{t}_2^{\,*} + \ldots + \widehat{t}_{1000}^{\,*}\right)/1000$
10: Change-point location estimation is given by $\widehat{t}^{\,*} = [L]$, the integer part of $L$

*Example.* We consider the model (1) for $\rho_0 = \rho_1 = 0$, $\theta_2 = 0$, $\delta_0(X_{t-1}) = \sqrt{0.04 + 0.36X_{t-1}^2}$, $\vartheta_1 = 1$, $\vartheta_2 = 1 + \phi$ and $\varepsilon_t \sim \mathcal{N}(0, 1)$. The resulting model is an ARCH(1). The change location estimators are calculated for $\phi = 0.3,\ 0.8$ and 1.5 at the locations $t^* = \tau \times n$ for $\tau = 0.25,\ 0.5$ and 0.75. In each case, we compute the bias and the standard error SE (SE=SD$/\sqrt{n}$), where SD denotes the standard deviation) of the change location estimator. Table 1 shows that the bias declines rapidly as $\phi$ increases. Also, as the sample size $n$ increases, the bias and the SE decrease. This tends to show the consistency of $\widehat{t}^{\,*}$, as expected from the asymptotic results.
We also consider the case $\varepsilon_t = \beta\varepsilon_{t-1} + \gamma_t$, where $|\beta| < 1$ and $\gamma_t \sim \mathcal{N}(0, \sqrt{1-\beta^2})$. It is easy to check that with this $(\varepsilon_t)_{t\in\mathbb{Z}}$ is stationary and strongly mixing, and that $\mathbb{E}(\varepsilon_t) = 0$ and $\mathbb{V}(\varepsilon_t) = 1$. In this case, we only study the SE for $n = 5000,\ 10000$ and the results are compared to those obtained for $\varepsilon_t \sim \mathcal{N}(0, 1)$, for the same values of $\phi$ as above but for $\tau = 0.25$ and 0.75. These results listed in Table 2 show that for $\varepsilon_t \sim \mathcal{N}(0, 1)$, the location estimator is more accurate and the SE decreases slightly compared to the case $\varepsilon_t \sim$ AR(1). It seems from these results that the nature of the white noise $\varepsilon_t$ does not much affect the location estimator for larger values of $n$ and $\phi$.

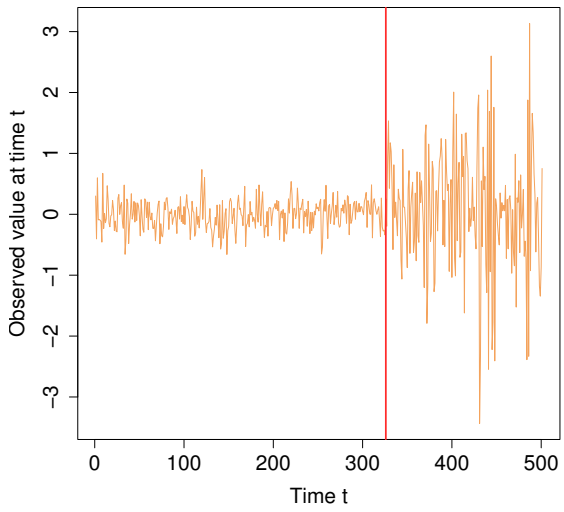
TABLE 1 – Change location estimation, its bias and SE for several values of $\phi$, $n$ and $\tau$ for iid $\varepsilon_t \sim \mathcal{N}(0,1)$.

| $\phi$ | $n$ | $t^* = 0.25n$ ($\tau = 0.25$) | | | $t^* = 0.5n$ ($\tau = 0.5$) | | | $t^* = 0.75n$ ($\tau = 0.75$) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\widehat{t}^*$ | SE | Bias | $\widehat{t}^*$ | SE | Bias | $\widehat{t}^*$ | SE | Bias |
| 0.3 | 500 | 181 | 4.9667 | 0.1120 | 277 | 3.4284 | 0.0540 | 384 | 3.3993 | 0.0180 |
| | 1000 | 287 | 3.8961 | 0.0370 | 522 | 2.4946 | 0.0220 | 767 | 2.8024 | 0.0170 |
| | 5000 | 1264 | 0.6270 | 0.0028 | 2516 | 0.6260 | 0.0032 | 3765 | 0.8172 | 0.0030 |
| | 10000 | 2517 | 0.4398 | 0.0017 | 5015 | 0.4131 | 0.0015 | 7515 | 0.4701 | 0.0015 |
| 0.8 | 500 | 137 | 1.8286 | 0.0240 | 258 | 0.9659 | 0.0160 | 383 | 1.0514 | 0.0160 |
| | 1000 | 257 | 0.5079 | 0.0070 | 507 | 0.6687 | 0.0070 | 757 | 0.6378 | 0.0070 |
| | 5000 | 1256 | 0.1874 | 0.0012 | 2506 | 0.1750 | 0.0012 | 3755 | 0.1602 | 0.0010 |
| | 10000 | 2506 | 0.1230 | 0.0006 | 5006 | 0.1388 | 0.0006 | 7505 | 0.1169 | 0.0005 |
| 1.5 | 500 | 130 | 0.8538 | 0.0100 | 254 | 0.4724 | 0.0080 | 379 | 0.4611 | 0.0080 |
| | 1000 | 253 | 0.2662 | 0.0030 | 504 | 0.2884 | 0.0040 | 753 | 0.2562 | 0.0030 |
| | 5000 | 1254 | 0.1344 | 0.0008 | 2503 | 0.1053 | 0.0006 | 3754 | 0.1174 | 0.0008 |
| | 10000 | 2504 | 0.0880 | 0.0004 | 5004 | 0.0842 | 0.0004 | 7504 | 0.0753 | 0.0004 |

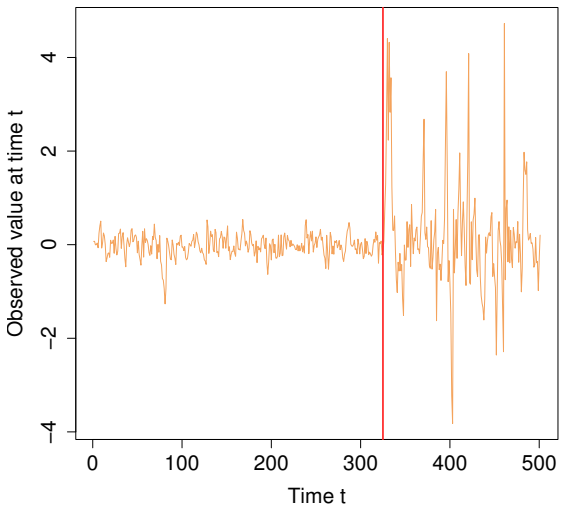
TABLE 2 – Change location estimation, its bias and SE for several values of $\phi$, $n$ and $\tau$ for iid $\varepsilon_t \sim \mathcal{N}(0,1)$ and for $\varepsilon_t \sim$ AR(1).

| $\phi$ | $n$ | $t^* = 0.25n$ ($\tau = 0.25$) | | | | | | $t^* = 0.75n$ ($\tau = 0.75$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\varepsilon_t \sim \mathcal{N}(0,1)$ | | | $\varepsilon_t \sim$ AR(1) | | | $\varepsilon_t \sim \mathcal{N}(0,1)$ | | | $\varepsilon_t \sim$ AR(1) | | |
| | | $\widehat{t}^*$ | SE | Bias | $\widehat{t}^*$ | SE | Bias | $\widehat{t}^*$ | SE | Bias | $\widehat{t}^*$ | SE | Bias |
| 0.3 | 5000 | 1265 | 0.6091 | 0.0030 | 1286 | 2.6225 | 0.0072 | 3766 | 0.6596 | 0.0032 | 3776 | 1.2243 | 0.0052 |
| | 10000 | 2516 | 0.4471 | 0.0016 | 2525 | 0.7136 | 0.0025 | 7515 | 0.4182 | 0.0015 | 7523 | 0.6563 | 0.0023 |
| 0.8 | 5000 | 1256 | 0.1701 | 0.0012 | 1260 | 0.2760 | 0.0020 | 3756 | 0.1818 | 0.0012 | 3760 | 0.2870 | 0.0020 |
| | 10000 | 2506 | 0.1482 | 0.0006 | 2510 | 0.1907 | 0.0010 | 7506 | 0.1338 | 0.0006 | 7509 | 0.1835 | 0.0009 |
| 1.5 | 5000 | 1254 | 0.1165 | 0.0008 | 1256 | 0.1835 | 0.0012 | 3754 | 0.1154 | 0.0008 | 3756 | 0.1784 | 0.0012 |
| | 10000 | 2503 | 0.0807 | 0.0003 | 2506 | 0.1284 | 0.0006 | 7504 | 0.0776 | 0.0004 | 7506 | 0.1263 | 0.0006 |

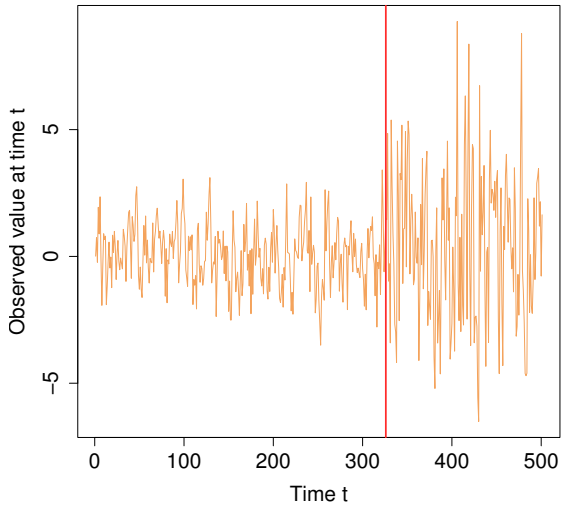
We present two graphs showing a change in volatility at a time $\widehat{t}^*$. This is indicated by a vertical red line on both graphics where one can easily see the evolution of the time series before and after the change location estimator $\widehat{t}^*$. The series in both figures are obtained for $m(\rho; x) = 0$, $\delta_0(x) = \sqrt{1 + 0.036x^2}$, $n = 500$, $\tau = 0.65$, and $\phi = 0.8$. That in Figure 1a is obtained for standard iid Gaussian $\varepsilon_t$s. In this case, using our method, the change location $t^* = 0.65 \times 500 = 325$ is estimated by $\widehat{t}^* = 326$. The time series in Figure 1b is obtained with $\varepsilon_t \sim$ AR(1). In this case, $t^*$ is estimated by $\widehat{t}^* = 325$.



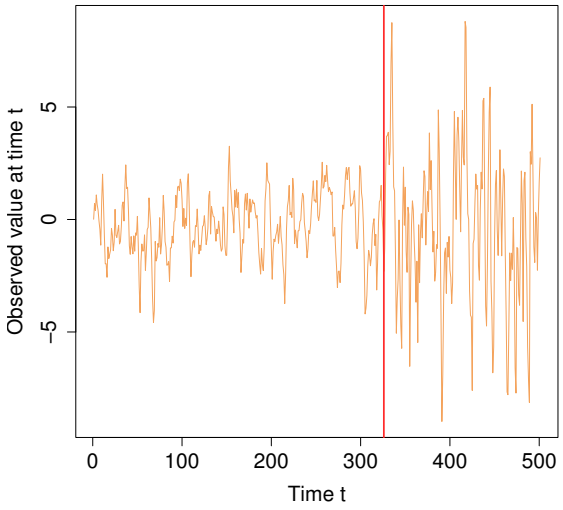

(**a**) ARCH(1) model, $\widehat{t}^* = 326$




(**b**) ARCH(1) model, $\widehat{t}^* = 325$




(**c**) CHARN model, $\widehat{t}^* = 326$




(**d**) CHARN model, $\widehat{t}^* = 326$

FIGURE 1 – Estimation of change-point in volatility for 500 observations. (**a**) ARCH(1) model with change point at $\widehat{t}^* = 326$; (**b**) ARCH(1) model with change point at $\widehat{t}^* = 325$; (**c**) CHARN model with change point at $\widehat{t}^* = 326$; (**d**) CHARN model with change point at $\widehat{t}^* = 326$.

## 3.1 Comparison with Some Recent Algorithms

We compare our method, referred to as LS, with the Wild Binary Segmentation (WBS) method studied in Fryzlewicz (2014), and one of its variants, called Narrowest-Over-Threshold (NOT), proposed by Baranowski et al. (2019), as well as, the Iterative Cumulative Sum

of Squares (ICSS) algorithm suggested by Inclan and Tiao (1994). All these methods are implemented under R software, and can, respectively, be found in the packages *wbs*, *not* and *ICSS*. Our comparison is based on $n$ observations simulated from (1) for $\rho_0 = \rho_1 = 0$, $\theta_2 = 0$, $\delta_0(X_{t-1}) = \sqrt{0.04 + 0.36X_{t-1}^2}$, $\vartheta_1 = 1$, $\vartheta_2 = 1 + \phi$ and $\varepsilon_t \sim \mathcal{N}(0,1)$. The change location estimators are calculated for $\phi = 0.3,\ 0.8$ and 1.5 at the locations $t^* = \tau \times n$ for $\tau = 0.25$ and 0.75.
From the results obtained (see Table 3), WBS, NOT and ICSS generate sequences of change-point location estimates, some of which have values close to the true locations. For $n = 100$ and $n = 200$, LS generally provides more accurate estimates $\widehat{t^*}$ of the true change-point location than WBS, NOT and ICSS, for different $\phi$ and locations $t^*$. Among all, it is generally the best, especially for larger values of $n$ and $\phi$.

TABLE 3 – Estimates of change location derived from LS, WBS, NOT and ICSS for a sample with a single break.

| | | $\boldsymbol{\phi = 0.3}$ | | $\boldsymbol{\phi = 0.8}$ | | $\boldsymbol{\phi = 1.5}$ | |
|---|---|---|---|---|---|---|---|
| | | $\boldsymbol{t^* = 0.25n}$ | $\boldsymbol{t^* = 0.75n}$ | $\boldsymbol{t^* = 0.25n}$ | $\boldsymbol{t^* = 0.75n}$ | $\boldsymbol{t^* = 0.25n}$ | $\boldsymbol{t^* = 0.75n}$ |
| **Methods** | $\boldsymbol{n}$ | $\boldsymbol{\widehat{t^*}}$ | $\boldsymbol{\widehat{t^*}}$ | $\boldsymbol{\widehat{t^*}}$ | $\boldsymbol{\widehat{t^*}}$ | $\boldsymbol{\widehat{t^*}}$ | $\boldsymbol{\widehat{t^*}}$ |
| LS | 100 | 39 | 70 | 31 | 79 | 28 | 78 |
| WBS | | 46\|43 | 87\|90 | 38\|39 | 82\|80 | 29\|30 | 78\|76 |
| NOT | | 51\|55 | 96 | 38\|42 | 79\|82\|91\|94 | 38\|35 | 78\|82 |
| ICSS | | 42\|48 | 68 | 31\|43 | 78\|81 | 33\|41 | 78\|97 |
| LS | 200 | 57 | 145 | 54 | 157 | 52 | 155 |
| WBS | | 70\|66 | 175\|177 | 62\|59 | 171\|167 | 58\|57 | 159\|158 |
| NOT | | 66\|70\|77 | 175\|190\|194 | 67\|70 | 158\|161 | 64\|68 | 155\|162 |
| ICSS | | 66\|87 | 80\|159 | 61\|76\|87 | 159 | 59\|66\|170\|187 | 156\|173 |

## 3.2 Application to Real Data

In this section, we apply our procedure to two sets of genuine time series, namely, the USA stock market prices. These data of length 2022 from the American stock market were recorded daily from 2 January 1992 to 31 December 1999. They represent the daily stock prices of the S&P 500 stock market (SPX). They are among the most closely followed stock market indices in the world and are considered as an indicator of the USA economy. They have also been recently examined by Kouamo et al. (2010) and can be found at `www.investing.com`. In Figure 2, we observe that the trend of the SPX daily stock price series is not constant over time. We also observe that stock prices have fallen sharply, especially in the time interval between 1997 and 1998.

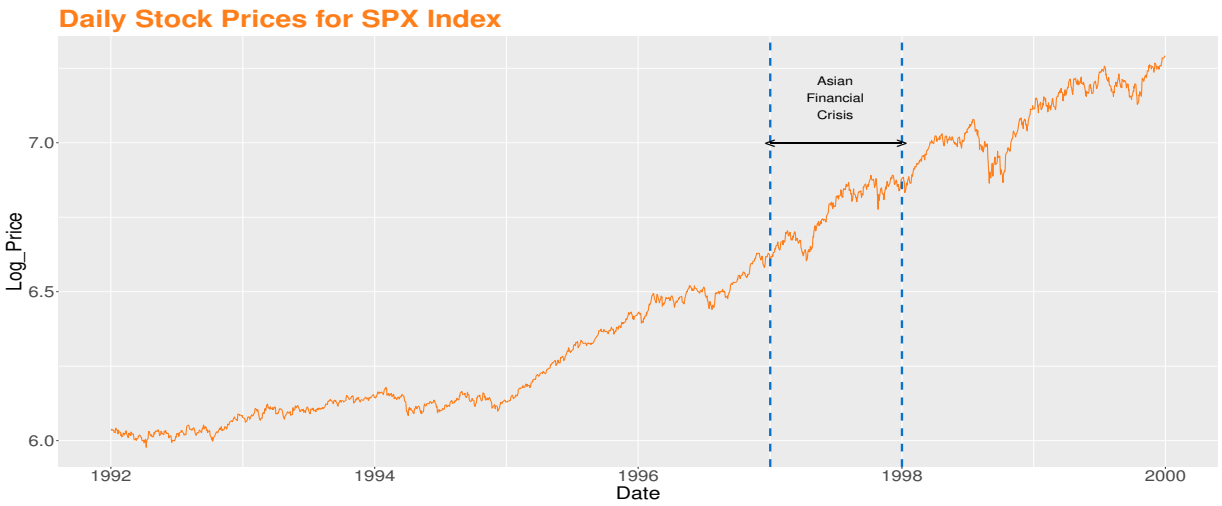



FIGURE 2 – Logarithmic series of S&P 500 stock prices from January 1992 to December 1999.

Denote by $D_t$ the value of the stock price for the SPX index at day $t$, and the first difference of the logarithm of stock price, $X_t$ as : $X_t = \log(D_t) - \log(D_{t-1}) = \log\left(\frac{D_t}{D_{t-1}}\right)$.
$X_t$ is the logarithmic return of stock price for the SPX index at day $t$.
The series $(X_t)$ is approximately piece-wise stationary on two segments and symmetric around zero (see Figure 3). This brought us to consider a CHARN model with $m(\rho; x) = 0$, $\delta_0(x) = \sqrt{\theta_0 + \theta_1 x^2}$ for $\theta_0 = 1$, $\theta_1 = 0$, $\vartheta_1$ and $\vartheta_2$ estimated by CLS described in Section 2.1. Using our procedure, we found an important change point in stock price volatility on 26 March 1997, which is consistent with the date found by Kouamo et al. (2010) (see Figure 3).

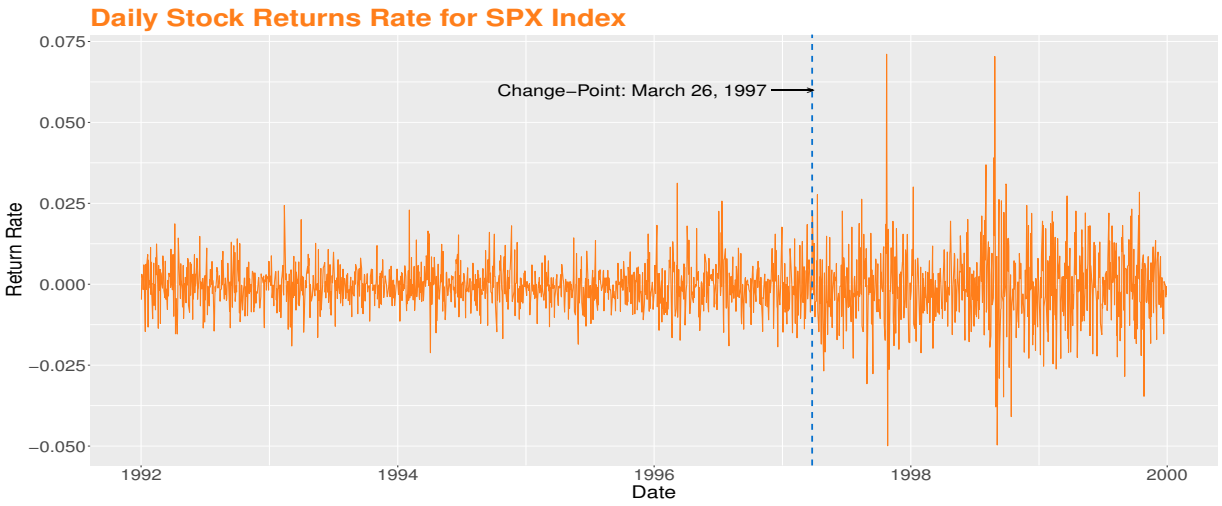



FIGURE 3 – Location of the change point in the volatility of the logarithmic stock price return series of the SPX Index from January 1992 to December 1999.

It should be noted that the change in volatility coincides with the Asian crisis in 1997 when

Thailand devalued its currency, the baht, against the US dollar. This decision led to a fall in the currencies and financial markets of several countries in its surroundings. The crisis then spread to other emerging countries with important social and political consequences and repercussions on the world economy.

# Bibliography


Bai, J. (1994), Least squares estimation of a shift in linear processes, *Journal of Time Series Analysis*, 15, pp. 453-472.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2019), *Narrowest-over-Threshold Change-Point Detection*, R package version 1, CRAN : Vienna, Austria.

Chen, J. and Gupta, A. (1999), Change point analysis of a Gaussian model, *Statistical Papers*, 40, pp. 323-333.

Csörgő, M. and Horváth, L. (1987), Nonparametric tests for the changepoint problem, *Journal of Statistical Planning and Inference*, 17, pp. 1-9.

Fryzlewicz, P. (2014), Wild Binary Segmentation for multiple change-point detection, *Annals of Statistics*, 42, pp. 2243-2281.

Hawkins, D.M. (1977), Testing a sequence of observations for a shift in location, *Journal of American Statistical Association*, 72, pp. 180-186.

Härdle, W. and Tsybakov, A. (1997), Local polynomial estimators of the volatility function in nonparametric autoregression, *Journal of Econometrics*, 81, pp. 223-242.

Härdle, W., Tsybakov, A. and Yang, L. (1998), Nonparametric vector autoregression, *Journal of Statistical Planning and Inference*, 68, pp. 221-245.

Hinkley, D.V. (1970), Inference about the change-point in a sequence of random variables, *Biometrika*, 57, pp. 1-17.

Hinkley, D.V. (1972), Time-ordered classification, *Biometrika*, 59, pp. 509-523.

Inclan, C. and Tiao, G.C. (1994), Use of cumulative sums of squares for retrospective detection of changes of variance, *Journal of American Statistical Association*, 89, pp. 913-923.

Kouamo, O., Moulines, E. and Roueff, F. (2010), Testing for homogeneity of variance in the wavelet domain, *Dependence in Probability and Statistics*, 200, pp. 175-205.

Ngatchou-Wandji, J. (2005), Checking nonlinear heteroscedastic time series models, *Journal of Statistical Planning and Inference*, 133, pp. 33-68.

Taniguchi, M. and Kakizawa, Y. (2000), *Asymptotic theory of estimation and testing for stochastic processes*, In Asymptotic Theory of Statistical Inference for Time Series, Springer : New York, NY, USA, pp. 51-165.

Yao, Y.C. and Davis, R.A. (1986), The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates, *Sankhyā : The Indian Journal of Statistics, Series A*, pp. 339-353.

# Filtre de Kalman Robuste avec covariables stochastiques

Jean-Luc MAHOROMEZA [1,2] & Olivier WINTENBERGER [1] & Adeline FERMANIAN [2] & Joseph DE VILMAREST [3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France*
[2] *LOPF, Califrais' Machine Learning Lab, Paris, France*
[3] *Viking Conseil, Paris, France*

**Résumé.** Le filtre de Kalman-Bucy permet l'estimation et la prédiction des modèles espace-état. Il est souvent confronté à des observations incertaines, par exemple quand elles proviennent de mesures bruitées d'un signal physique (capteurs, météo, etc.). Plutôt que de considérer les covariables comme des quantités déterministes, ce qui est fait classiquement, notre objectif est de prendre en compte leurs incertitudes et d'étudier la robustesse du filtre. Plus précisément, le filtre de Kalman est une approche bayésienne qui donne l'estimation de la loi *a posteriori* de l'état, conditionnellement aux observations passées. Nous raffinons cette approche en intégrant les incertitude des covariables dans le modèle, c'est-à-dire en les considérant comme des variables aléatoires dont la loi doit être estimée.

**Mots-clés.** Robustesse, Séries temporelles, Filtre de Kalman, Statistiques bayésiennes.

**Abstract.** The Kalman-Bucy filter is used for estimating and predicting state-space models. It frequently faces uncertain observations, such as those derived from noisy measurements of physical signals (including sensors, weather data, etc.). Instead of considering covariates as deterministic quantities, as is conventionally done, our purpose is to consider their uncertainties into account and address a robust filter. More specifically, the Kalman filter is a Bayesian approach that estimates the posterior distribution of the state conditionally on past observations. We refine this approach by integrating the uncertainties of the covariates into the model, i.e, considering them as stochastic variables whose distributions need to be estimated.

**Keywords.** Robustness, Time series, Kalman Filter, Bayesian statistics.

## 1 Introduction

Les modèles espace-état se sont imposés comme un cadre incontournable pour analyser les modèles dynamiques. Ces modèles reposent sur l'idée qu'un vecteur observable $y_t$, souvent appelé "vecteur d'observation", est une fonction du vecteur non observé (latent) $\theta_t$, désigné comme le "vecteur d'état". La dynamique régissant le vecteur d'observation et le vecteur d'état est encapsulée au sein du modèle espace-état. Dans la seconde moitié du 20ème siècle,

le filtre de Kalman and Bucy (1961) a considérablement avancé la modélisation des modèles espace-état. Il a fourni un cadre d'estimation des modèles linéaires, fondé sur le filtre de Wiener (Wiener, 1949). Le filtre de Kalman-Bucy se distingue par son mécanisme récursif pour estimer et prédire l'état des modèles dynamiques linéaires dans un contexte de données bruitées. Cet avancement a été crucial pour la prédiction précise (l'espérance conditionnelle de l'état sachant le passé est calculée de façon exacte), l'estimation et le contrôle dans les modèles dynamiques linéaires, marquant une réalisation significative dans l'ingénierie, y compris le suivi de cibles, le contrôle d'avions (Ray and Stengel, 1991), l'ingénierie électrique (Vilmarest, 2022) et les sciences environnementales.

Au fil du temps, le filtre de Kalman a connu un succès généralisé, menant à diverses interprétations. Ollivier (2018) a établi un lien entre le filtre de Kalman étendu (Fahrmeir, 1992) et le gradient naturel en ligne de Murata and Amari (1999). De plus, Durbin and Koopman (2012) définit le problème de filtrage dans une configuration gaussienne linéaire comme un problème de recherche d'un estimateur linéaire non biaisé de variance minimale. Plusieurs variantes ont été développées pour adapter le filtre de Kalman à une gamme de problèmes. Ces adaptations ont exploré de nombreuses directions, incluant : (1) l'extension du filtrage aux scénarios non linéaires (Wan and Van Der Merwe, 2000; Fahrmeir, 1992); (2) le raffinement du filtre pour les cas où les hyper-paramètres du modèle sont inconnus (de Vilmarest and Wintenberger, 2021); (3) l'amélioration de la robustesse du filtre face aux perturbations du modèle, incluant la prise en compte à la fois du bruit statistiquement inconnu au sein du modèle espace-état et des incertitudes dans les variables du modèle dues à des erreurs déterministes ou stochastiques (Yedavalli, 2014; Zeng et al., 2012; Ra and Whang, 2008; Theodor and Shaked, 1996; Stengel and Ray, 1991; Ray and Stengel, 1991); et (4) la prise en compte des erreurs de modélisation où le modèle idéal est souvent mal défini, nécessitant parfois des hypothèses trop fortes qui peuvent s'écarter nettement des conditions réelles (Toda and Patel, 1980; Nishimura, 1970).

Notre étude aborde le problème (3), se concentrant sur la construction d'un estimateur robuste aux perturbations qui se manifestent comme du bruit dans les covariables. De même, notre étude s'inscrit dans la direction du problème (4) ; l'hypothèse courante que les covariables sont déterministes néglige leur nature intrinsèquement bruitée ou estimée. L'étude de la robustesse d'un estimateur se divise en deux grandes catégories : la première suppose une perturbation déterministe, sans aucun modèle probabiliste attribué à cette perturbation (Yedavalli, 2014). La seconde catégorie couvre les systèmes soumis à des perturbations stochastiques (Ra and Whang, 2008; Theodor and Shaked, 1996; Stengel and Ray, 1991).

Dans notre étude, nous considérons un modèle bien spécifié mais dans lequel les covariables sont bruitées. Ce cas des covariables soumises aux perturbations stochastiques se rencontre dans un large spectre d'applications. Par exemple, dans les sciences environnementales, elle est essentielle pour modéliser la dispersion des polluants et les prévisions du changement climatique, où les covariables comme les concentrations de polluants ou les mesures de température viennent avec leur propre ensemble d'incertitudes. Le secteur financier s'appuie sur des méthodologies similaires pour modéliser les indicateurs économiques et les prix des actifs, en intégrant la volatilité et les évaluations des risques pour prendre des décisions d'investissement éclairées. Dans le domaine de la santé, en particulier dans le

domaine de l'épidémiologie, la modélisation de la propagation des maladies avec des données incertaines est critique pour les stratégies de planification et de réponse. De plus, l'ingénierie, en particulier les systèmes de contrôle et la robotique, applique ces principes pour compenser les imprécisions des capteurs et pour affiner les réponses du système. La météorologie se démarque également, où la précision des prédictions météorologiques est considérablement améliorée en modélisant la perturbation dans les données atmosphériques.

# 2 Énoncé du problème

Nous explorons le modèle espace-état linéaire gaussien défini comme suit :

$$\begin{aligned} &\text{Équation d'état :} && \theta_t = A\theta_{t-1} + \eta_t, && \eta_t \sim \mathcal{N}(0, Q_t), \\ &\text{Équation des covariables :} && x_t = \tilde{x}_t + \Delta x_t, && \Delta x_t \sim \mathcal{N}(0, R_t), \\ &\text{Équation d'espace :} && y_t = x_t^T \theta_t + \varepsilon_t, && \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2). \end{aligned} \tag{1}$$

Ici, $y_t$ et $\varepsilon_t$ sont des nombres réels, tandis que $\theta_t$, $x_t$, $\tilde{x}_t$, $\Delta x_t$ et $\eta_t$ sont des vecteurs dans $\mathbb{R}^d$. La transition de $\theta_{t-1}$ à $\theta_t$ est donnée par une matrice de transition $A \in \mathbb{R}^{d\times d}$ supposée connue, plus un bruit gaussien $\eta_t$ centré de matrice de covariance $Q_t \in \mathbb{R}^{d\times d}$. L'équation d'espace régit la relation entre l'observation $y_t$, les variables explicatives $x_t$ et l'état $\theta_t$. Les hyperparamètres du modèle, $Q_t$, $R_t$ et $\sigma_t^2$, sont supposés être connus. Nous supposons que $\Delta x_t$ suit une loi gaussienne centrée de matrice de covariance $R_t$.

Désignons par $\mathcal{F}_m = \sigma(x_1, y_1, \ldots, x_m, y_m)$ la filtration canonique représentant l'information disponible jusqu'au temps $m$. On se concentre sur les scénarios où le modèle devient incomplet ; c'est-à-dire que les valeurs de $x_t$ manquent à partir d'un certain temps $m$, et seul $\tilde{x}_t$ est observé. Nous souhaitons donner l'expression des moments d'ordre un et deux de la loi *a posteriori* de l'état. Formellement, nous exprimons ces quantités de la manière suivante :

$$\begin{aligned} \tilde{\theta}_{t|m} &= \mathbb{E}[\theta_t | \mathcal{F}_{m-1}, y_m], \\ \tilde{P}_{t|m} &= \mathbb{E}\big[(\theta_t - \tilde{\theta}_{t|m})(\theta_t - \tilde{\theta}_{t|m})^T | \mathcal{F}_{m-1}, y_m\big]. \end{aligned} \tag{2}$$

# 3 Présentation de l'approche

Commençons par rappeler l'expression du filtre de Kalman standard où les covariables $x_t$ sont observées à chaque instant. Dans le cas du modèle espace-état linéaire gaussien, l'espérance de $\theta_t$ conditionnellement à la filtration $\mathcal{F}_t$ est une gaussienne de moyenne $\hat{\theta}_{t|t}$ et de matrice de covariance $P_{t|t}$ (Kalman and Bucy, 1961). Les quantités $\hat{\theta}_{t|t}$ et $P_{t|t}$ sont données dans le théorème 3.1.

**Théorème 3.1 (Filtre de Kalman)** *Soit $t > 0$, sous les hypothèses du modèle espace-état (1), la loi de $\theta_t$, conditionnellement à la filtration $\mathcal{F}_t$, $\theta_t$ suit une gaussienne de moyenne $\hat{\theta}_{t|t}$*

*et de matrice de covariance $P_{t|t}$ dont les expressions sont :*

$$P_{t|t} = P_{t|t-1} - K_t x_t^T P_{t|t-1}, \qquad P_{t+1|t} = \Phi P_{t|t} \Phi^T + Q_t,$$
$$\hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} - K_t (y_t - x_t^T \hat{\theta}_{t|t-1}), \qquad \hat{\theta}_{t+1|t} = \Phi \hat{\theta}_{t|t},$$

*où*

$$K_t = \frac{P_{t|t-1} x_t}{x_t^T P_{t|t-1} x_t + \sigma_t^2}.$$

Ici, $K_t$ est une matrice $\mathbb{R}^{d \times 1}$, appelée **matrice de gain**. Elle combine de manière optimale les nouvelles observations avec les estimations antérieures pour mettre à jour les prédictions de l'état en fonction de l'erreur en $y_t$. Utiliser $\tilde{x}_t$ dans le théorème 3.1 induit des écarts qui deviennent plus grandes en fonction de la variance de la perturbation $\Delta x_t$ (Figure 1).

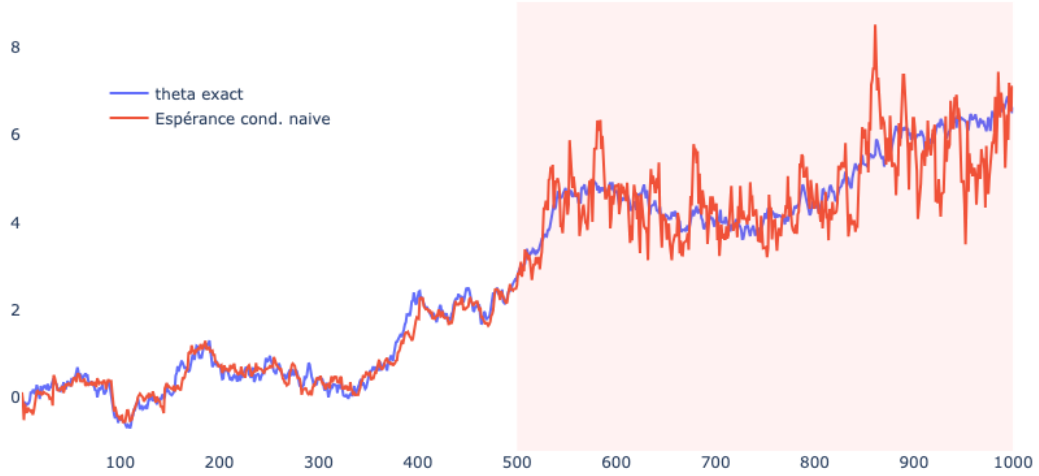



Figure 1: Simulation avec $\tilde{x}_t = \sin(10\pi t)$, $\Delta x_t \sim \mathcal{N}(0, 0.1)$, $\sigma_t^2 = 0.1$, $Q_t = 0.01$, $\theta_{0|0} = 1$, $P_{0|0} = 1$ et $A = I$. La valeur exacte de $\theta_t$ est représentée en bleu, la courbe rouge représente la moyenne $\hat{\theta}_{t|t}$. $x_t$ est observé jusqu'à $t = 500$. On observe qu'à partir $t = 500$, le filtre devie.

Dans le cas où $x_t$ n'est pas observé, en conditionnant l'espérance conditionnelle $\mathbb{E}[\theta_t | \mathcal{F}_{t-1}, y_t]$ par $x_t$ (tower property), cela nous permet de reformuler l'espérance sous $\theta_t$ définie dans l'équation (2) en une espérance de $\hat{\theta}_{t|t}$ (Théorème 3.1) sous la loi *a posteriori* de $x_t$. Par la suite, on notera par $f_{x_t|\mathcal{F}_{t-1}, y_t}$, la densité de probabilité de la loi *a posteriori* de $x_t$. Notre

démarche consiste d'abord à donner l'expression exacte de la densité de probabilité $f_{x_t|\mathcal{F}_{t-1},y_t}$ à une constante de normalisation près. Cela nous permet par la suite de déterminer l'espérance de $\hat{\theta}_{t|t}$ sous la loi *a posteriori* de $x_t$, et par conséquent $\mathbb{E}[\theta_t|\mathcal{F}_{t-1}, y_t]$. Cette méthode offre un double avantage, car elle permet de filtrer à la fois le bruit sur $x_t$ et sur $\theta_t$ grâce à la nouvelle observation $y_t$. Cet avantage est particulièrement pertinent lorsque la variance de $\Delta x_t$ est élevée, comme le montre la Figure 2.

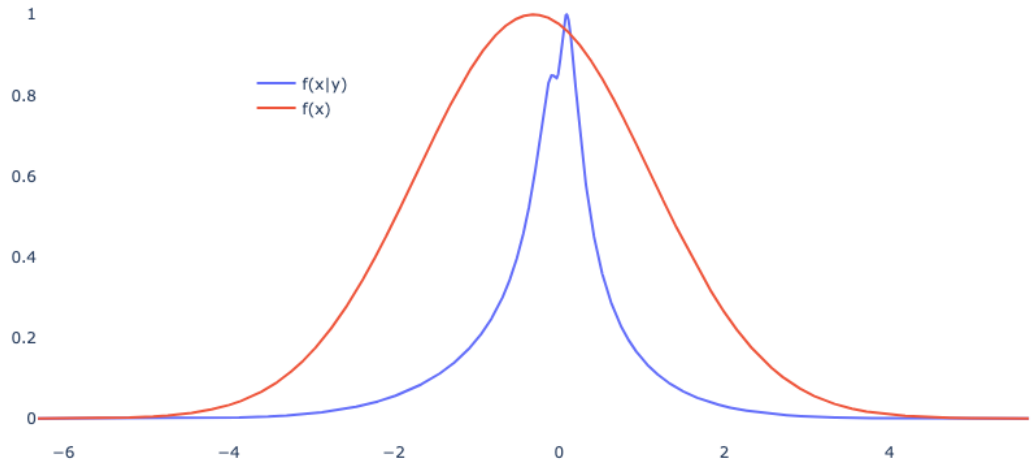



Figure 2: Simulation avec $\tilde{x}_t = \sin(10\pi t)$, $\Delta x_t \sim \mathcal{N}(0, 2)$, $\sigma_t^2 = 0.1$, $Q_t = 5$, $\theta_{0|0} = 1$, $P_{0|0} = 1$ et $A = I$. À $t$ fixé et à une constante de normalisation près, le graphe rouge représente la densité $f_{x_t} \sim \mathcal{N}(\tilde{x}_t, R_t)$ et le graphe bleu représente la densité de probabilité $f_{x_t|\mathcal{F}_{t-1},y_t}$.

# 4 Résultats théoriques

Commençons par considérer que l'instant $t$ représente la première fois que nous n'observons pas $x_t$ (c'est à dire $t = m$). Pour simplifier les équations, nous considérons que la matrice de transition $A$ est une matrice identité. Puisqu'à $t - 1$, nous avons observé $x_{t-1}$, l'espérance conditionnelle $\hat{\theta}_{t|t-1} = \mathbb{E}[\theta_t|\mathcal{F}_{t-1}]$ et sa matrice de covariance $P_{t|t-1} = \mathbb{E}[(\theta_t - \hat{\theta}_{t|t-1})(\theta_t - \hat{\theta}_{t|t-1})^T|\mathcal{F}_{t-1}]$ sont données de façon exacte par le filtre de Kalman. La densité de probabilité

$f_{\theta_t|\mathcal{F}_{t-1},y_t}$ de la loi *a posteriori* de $\theta_t$ se dérive comme suit :

$$\begin{aligned} f_{\theta_t|\mathcal{F}_{t-1},y_t}(z) &= \int_{\mathbb{R}^d} f_{\theta_t|\mathcal{F}_{t-1},y_t,x_t=x}(z) \cdot f_{x_t|\mathcal{F}_{t-1},y_t}(x)\, dx \\ &= \int_{\mathbb{R}^d} f_{\mathcal{N}(\hat{\theta}_{t|t},P_{t|t})}(z) \cdot f_{x_t|\mathcal{F}_{t-1},y_t}(x)\, dx. \end{aligned} \tag{3}$$

La densité de probabilité résultante est une loi de mélange non finie. Les lois de mélange ont l'avantage d'avoir une forme assez simple et explicite des leur moments.

**Hypothèse 4.1** *Soit $t > 0$, sous les hypothèses du modèle espace-état* (1)*, supposons que l'intégrale*

$$\int_{\mathbb{R}^d} x^2 f_{x_t|\mathcal{F}_{t-1},y_t}(x)\, dx$$

*est finie.*

**Proposition 4.1** *Soit $t > 0$, sous l'hypothèse* (4.1) *et du modèle espace-état* (1)*, les moments d'ordre un ($\tilde{\theta}_{t|t}$) et deux ($\tilde{P}_{t|m}$) de le loi a posteriori de $\theta_t$ sont données par les expressions :*

$$\tilde{\theta}_{t|t} = \mathbb{E}[\theta_t|\mathcal{F}_{t-1}, y_t] = \int_{\mathbb{R}^d} \hat{\theta}_{t|t} f_{x_t|\mathcal{F}_{t-1},y_t}(x)\, dx = \mathbb{E}_{x_t|\mathcal{F}_{t-1},y_t}[\hat{\theta}_{t|t}],$$

$$\tilde{P}_{t|t} + \left(\tilde{\theta}_{t|m}\right)\left(\tilde{\theta}_{t|m}\right)^T = \int_{\mathbb{R}^d} \left(P_{t|t} + \hat{\theta}_{t|t}\left(\hat{\theta}_{t|t}\right)^T\right) f_{x_t|\mathcal{F}_{t-1},y_t}(x)\, dx = \mathbb{E}_{x_t|\mathcal{F}_{t-1},y_t}[P_{t|t} + \hat{\theta}_{t|t}\left(\hat{\theta}_{t|t}\right)^T].$$

La proposition 4.1 est intéressante, car elle reformule le calcul des moments *a posteriori* de $\theta_t$, spécifiquement l'espérance et la variance, en termes d'espérances de fonctions dépendant de $x_t$ sous la loi $f_{x_t|\mathcal{F}_{t-1},y_t}$. L'évaluation de ces espérances nécessite la connaissance de l'expression de $f_{x_t|\mathcal{F}_{t-1},y_t}$, à une constante de normalisation près, que nous dérivons en appliquant la règle de Bayes.

**Proposition 4.2** *Soit $t > 0$, sous les hypothèses du modèle espace-état (1), à une constante de normalisation près, la densité de probabilité $f_{x_t|\mathcal{F}_{t-1},y_t}$ a pour expression :*

$$f_{x_t|\mathcal{F}_{t-1},y_t} \propto \frac{(2\pi)^{-(d+1)/2}|R_t|^{-1/2}}{(x_t^T P_{t|t-1} x_t + \sigma_t^2)^{1/2}} \exp(C) \exp\left(-\frac{1}{2}\left[(x_t - \Sigma_t K_t)^T \Sigma_t^{-1}(x_t - \Sigma_t K_t)\right]\right),$$

*où*

$$C = -\frac{1}{2}\left(\frac{y_t^2}{x_t^T P_{t|t-1} x_t + \sigma_t^2} + \tilde{x}_t^T R_t^{-1} \tilde{x}_t - K_t^T \Sigma_t^T K_t\right), \quad \Sigma_t = \left(\frac{\hat{\theta}_{t|t-1}\hat{\theta}_{t|t-1}^T}{x_t^T P_{t|t-1} x_t + \sigma_t^2} + R_t^{-1}\right)^{-1},$$

$$K_t = \frac{\hat{\theta}_{t|t-1} y_t}{x_t^T P_{t|t-1} x_t + \sigma_t^2} + R_t^{-1}\tilde{x}_t.$$

La connaissance de l'expression de la densité $f_{x_t|\mathcal{F}_{t-1},y_t}$ à une constante de normalisation près nous permet de construire un estimateur consistant de $\mathbb{E}_{x_t|\mathcal{F}_{t-1},y_t}[\hat{\theta}_{t|t}]$, qui est lui-même un estimateur de $\mathbb{E}[\theta_t|\mathcal{F}_{t-1}, y_t]$. À chaque $t > 0$ fixé, les méthodes de Monte-Carlo nous permettent de construire un estimateur consistant de $\mathbb{E}_{x_t|\mathcal{F}_{t-1},y_t}[\hat{\theta}_{t|t}]$ dont les bornes de l'erreur sont en $\mathcal{O}(n^{-1/2})$ où $n$ est la taille de l'échantillon $\{x_i\}_{0\leq i\leq n}$ distribué selon la loi $f_{x_t|\mathcal{F}_{t-1},y_t}$. La borne de l'erreur en $\mathcal{O}(n^{-1/2})$ peut-être améliorée en utilisant les méthode Quasi-Monte Carlo ou les méthodes de Monte-Carlo par chaînes de Markov. Les espressions des moments d'ordre un et deux dans le théorème 4.1 nous permettent d'optimiser le temps de calcul dû à l'échantillonnage en utilisant un même échantillon de $f_{x_t|\mathcal{F}_{t-1},y_t}$.

# References


Joseph de Vilmarest and Olivier Wintenberger. Viking: Variational bayesian variance tracking. *arXiv preprint arXiv:2104.10777*, 2021.

James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*, volume 38. OUP Oxford, 2012.

Ludwig Fahrmeir. Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87 (418):501–509, 1992.

Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. 1961.

Noboru Murata and Shun-ichi Amari. Statistical analysis of learning dynamics. *Signal Processing*, 74(1):3–28, 1999.

T Nishimura. Modeling errors in kalman filters. Technical report, 1970.

Yann Ollivier. Online natural gradient as a kalman filter. 2018.

Won-Sang Ra and Ick-Ho Whang. Stochastic robust kalman filtering for linear time-varying systems with a multiplicative measurement noise. *IFAC Proceedings Volumes*, 41(2):12546–12551, 2008.

Laura Ryan Ray and Robert F Stengel. Application of stochastic robustness to aircraft control systems. *Journal of Guidance, Control, and Dynamics*, 14(6):1251–1259, 1991.

Robert F Stengel and Laura R Ray. Technical notes and correspondence: Stochastic robustness of linear time-invariant control systems. *NASA. Langley Research Center, Joint University Program for Air Transportation Research, 1990-1991*, 1991.

Yahali Theodor and Uri Shaked. Robust discrete-time minimum-variance filtering. *IEEE Transactions on Signal Processing*, 44(2):181–189, 1996.

M. Toda and R. Patel. Bounds on estimation errors of discrete-time filters under modeling uncertainty. *IEEE Transactions on Automatic Control*, 25(6):1115–1121, 1980. doi: 10.1109/TAC.1980.1102502.

Joseph de Vilmarest. *Modèles espace-état pour la prévision de séries temporelles. Application aux marchés électriques*. PhD thesis, Sorbonne université, 2022.

Eric A Wan and Rudolph Van Der Merwe. The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pages 153–158. Ieee, 2000.

Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949.

Rama K Yedavalli. Robust control of uncertain dynamic systems. *AMC*, 10:12, 2014.

Caibin Zeng, YangQuan Chen, and Qigui Yang. Robust controllability of interval fractional order linear time invariant stochastic systems. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 4047–4050, 2012. doi: 10.1109/CDC.2012.6425949.

# Données de composition, de distribution et d'échelle

# Wasserstein multivariate auto-regressive models for modeling distributional time series and its application in graph learning

Yiye Jiang [1]

*Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France*
*yiye.jiang@inria.fr*

**Résumé.** Nous proposons un nouveau modèle auto-régressif pour l'analyse statistique des séries chronologiques distribuées multivariées. Les données d'intérêt consistent en une collection de plusieurs séries de mesures de probabilité supportées sur un intervalle borné de la ligne réelle, et qui sont indexées par des instants de temps. Les mesures de probabilité sont modélisées comme des objets aléatoires dans un espace de Wasserstein. Nous établissons le modèle auto-régressif dans l'espace tangent à la mesure de Lebesgue en centrant d'abord toutes les mesures brutes de manière à ce que leur moyenne de Fréchet corresponde à la mesure de Lebesgue. En utilisant la théorie des systèmes de fonctions aléatoires itérés, des résultats sur l'existence, l'unicité et la stationnarité de la solution d'un tel modèle sont fournis. Nous proposons également un estimateur consistant pour le coefficient du modèle. En plus de l'analyse de données simulées, le modèle proposé est illustré avec deux ensembles de données réelles, constitués d'observations des distributions d'âge de différents pays / états pour l'un, et du réseau de vélos en libre-service à Paris pour l'autre. Enfin, grâce aux contraintes du simplexe que nous imposons sur les coefficients du modèle, l'estimateur proposé, qui est appris sous ces contraintes, a naturellement une structure peu dense. Cette structure permet en outre d'appliquer le modèle proposé à l'apprentissage d'un graphe de dépendance temporelle à partir de séries chronologiques distribuées multivariées.

**Mots-clés.** Modèles autorégressifs, espaces de Wasserstein, graph learning, analyse de données distribuées.

**Abstract.** We propose a new auto-regressive model for the statistical analysis of multivariate distributional time series. The data of interest consist of a collection of multiple series of probability measures supported over a bounded interval of the real line, and that are indexed by distinct time instants. The probability measures are modelled as random objects in a Wasserstein space. We establish the auto-regressive model in the tangent space at the Lebesgue measure by first centering all the raw measures so that their Fréchet means turn to be the Lebesgue measure. Using the theory of iterated random function systems, results on the existence, uniqueness and stationarity of the solution of such model are provided. We also propose a consistent estimator for the model coefficient. In addition to the analysis of simulated data, the proposed model is illustrated with two real data sets made of observations from age distribution in different countries / states and bike sharing network in Paris. Finally, due to the simplex constraints that we impose on the model coefficients, the proposed estimator that is learned under these constraints, naturally has a sparse structure. The sparsity allows furthermore the application of the proposed model in learning a graph of temporal dependency from the multivariate distributional time series.

**Keywords.** Autoregressive models, Wasserstein spaces, graph learning, distributional data analysis

## 1 Introduction

Distributional time series is a recent research field that deals with observations that can be modeled as sequences of time-dependent probability distributions. Such distributional time series are ubiquitous in many scientific fields. A pertinent example is the analysis of sequences of the indicator distributions supported over age intervals, such as mortality and fertility (Mazzuco and Scarpa, 2015; Shang and Haberman, 2020), over calendar years in demographic studies. In Figure 1, we illustrate the data type with a real data set.

[1]This work was conducted while the author was preparing her PhD at Université de Bordeaux.

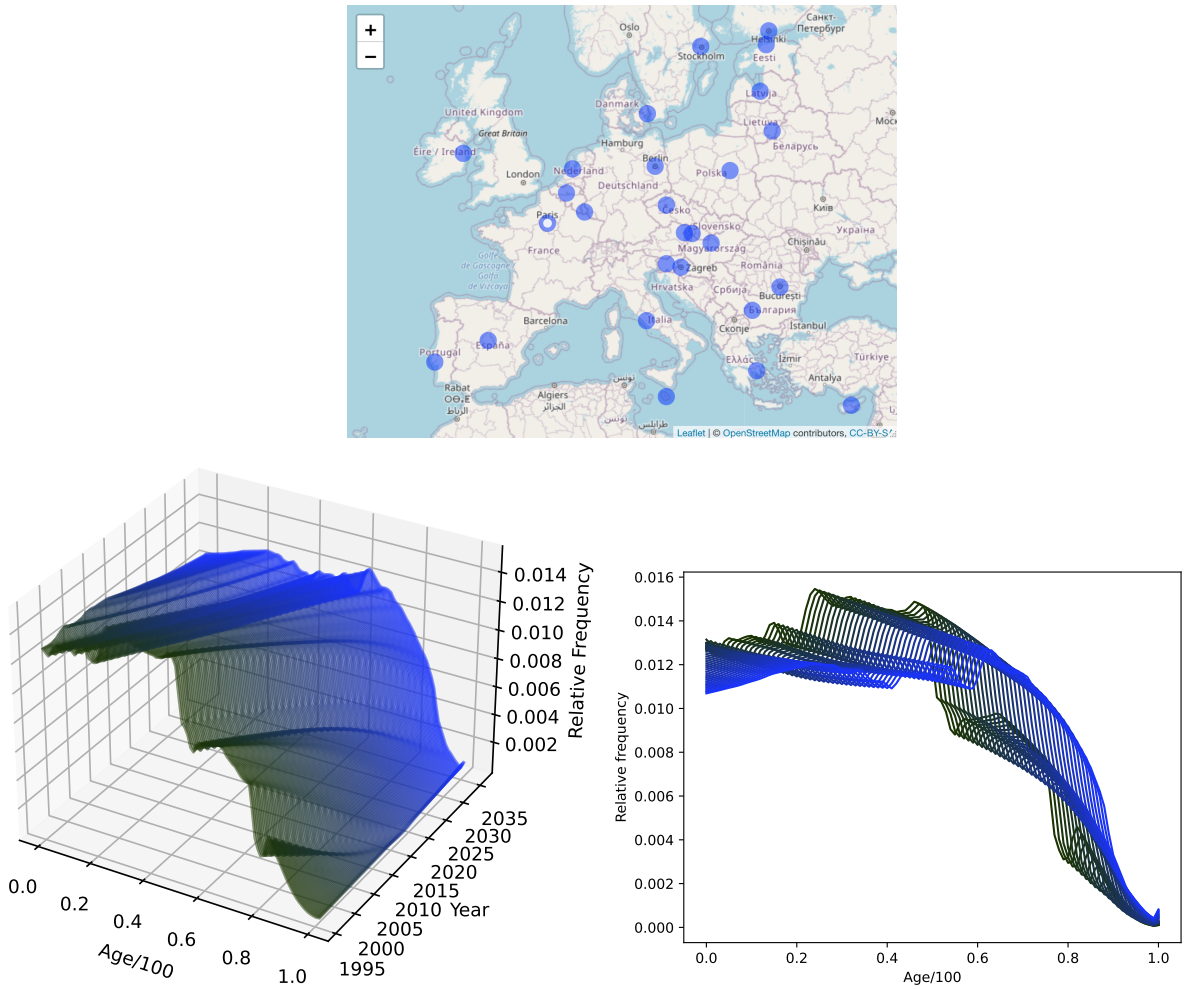



Figure 1: *Annual records of age distributions of countries/states.* On the top are 27 countries in the European union. A sequence of age distribution is recorded at each country over years. For example, at the bottom we illustrate the sequence of France, where one distribution supported over $[0, 1]$ is observed at each year. On the lower left, we visualize the resulting univariate distributional time series with a surface in the coordinate system of Age $\times$ Year $\times$ Relative frequency. The raw data in this plot consist in 40 annual distributions. We complete them with interpolated samples to draw the surface. On the lower right, we show the projection of the raw time series onto the Age $\times$ Relative frequency plane. We can see that the population is aging along time.

Since distributions can be characterized by functions, such as densities, quantile functions, and cumulative distribution functions, to analyze the distributional time series, one may turn to study one of its functional representations with the tools from functional time series analysis (Bosq, 2000). However, due to the nonlinear constraints, such as monotonicity and positivity, the representing functions of distributions do not constitute linear spaces. Consequently, basic notions needed by the classical tools, such as additivity and scalar multiplication, do not adapt, in a straightforward manner. This causes models devised for random elements of a Hilbert space often to fail. One existing approach is to map distributions to unconstrained functions by the log quantile density (LQD) transformation (Petersen and Müller, 2016), and then apply the functional tools (Kokoszka et al., 2019). However LQD does not take into account the geometry of the distribution space, thus it can lead to deformations in the distance. Recent approaches consider such geometry by adopting the Wasserstein metric (Bigot et al., 2017; Panaretos and Zemel, 2016; Petersen and Müller, 2019) of distributions. For further reading on the topic of statistics in Wasserstein space, we refer to Bigot (2020); Panaretos and Zemel (2020); Petersen et al. (2022).

Relying on the theories of Wasserstein spaces, Chen et al. (2021); Zhang et al. (2021); Zhu and Müller (2021) have successfully extended one of the most important models in classical time series analysis, autoregressive (AR) model, to univariate distributional (with closed bounded supports) time series, namely, $(\boldsymbol{\mu}_t)_{t\in\mathbb{Z}} \in \mathcal{P}([0,1])$. In this work, we will furthermore extend the AR model to the multivariate distributional (with closed bounded supports) case, that is, $(\boldsymbol{\mu}_t^i)_{t\in\mathbb{Z}}$ for $i = 1, \ldots, N \in \mathcal{P}([0,1])$, based on the tools of Wasserstein spaces as well. In the following, we firstly recall the classical AR models. Next we introduce the necessary notions of the Wasserstein spaces to present our model. Then we present our model and its theoretical properties. Fourthly, we propose a consistent estimator for the model coefficient. Lastly we fit our model over a real data sets, and show the numerical results.

# 2 Backgrounds

## 2.1 Vector autoregressive models of order 1

In this section, we recall the classical AR models in the multivariate setting. Let $\boldsymbol{x}_{it} \in \mathbb{R}$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$, a multivariate time series, and assume $\mathbb{E}\boldsymbol{x}_{it} = u_i$ exists and time invariant, then the vector autoregressive model of order 1 (VAR(1)) writes as

$$\boldsymbol{x}_{it} - u_i = \sum_{j=1}^{N} A_{ij}(\boldsymbol{x}_{j,t-1} - u_j) + \boldsymbol{\epsilon}_{it},$$

where $\boldsymbol{\epsilon}_{it} \sim WN(0, \sigma^2)$ with $\sigma$ non-zero, and $\epsilon_{it}, \epsilon_{jt}, i \neq j$ are not correlated. Then the goal of this work is to extend this model by "replacing" each scalar $\boldsymbol{x}_{it}$ by a univariate distribution $\mu_i^t \in \mathcal{P}(\mathbb{R})$. The main ingredient is using the geometrical notions of Wasserstein space. In the next section, we introduce the ones which are needed by our models.

## 2.2 Notions in Wasserstein spaces

We firstly define the Wasserstein spaces. Since we will need the notion of Tangent space, we use the spaces of 2-Wasserstein distances. More specifically, we focus on the space

$$\mathcal{W}_2(\mathbb{R}) = \left\{ \mu \in \mathcal{P}(\mathbb{R}) \middle| \int_{\mathbb{R}} x^2 d\mu(x) < \infty \right\}, \tag{2.1}$$

endowed with the 2-Wasserstein distance[2]

$$d_W(\mu, \nu) = \int_0^1 \left( F_\mu^{-1}(u) - F_\nu^{-1}(u) \right)^2 du, \tag{2.2}$$

[2]In general, the Wasserstein distances are defined by the Kantorovich problems. However, in the special case we consider here, that is, the cost is quadratic and the domain is $\mathbb{R}$, the solution of Kantorovich problem is explicit and given by the definition in Equation (2.2).

where $F_{\mu}^{-1}(u), F_{\nu}^{-1}(u)$ are the quantile functions of $\mu$ and $\nu$.

The space defined in Equation (2.1) is not linear. Nevertheless, Wasserstein spaces of order 2 permit the notion of Tangent spaces (Ambrosio et al., 2008; Bigot et al., 2017; Zemel and Panaretos, 2019), where the linear operations are enabled again. To present the Tangent spaces, we firstly recall the notion of pushforward for a pair of measures.

Given two measurable spaces $(\mathcal{X}_i, \sigma_i)$, $i = 1, 2$, a measurable function $f : \mathcal{X}_1 \to \mathcal{X}_2$, and a measure $\mu_1 : \sigma_1 \to [0, +\infty]$, the pushforward measure of $\mu_1$ by $f$ denoted by $f\#\mu_1$ is defined as

$$f\#\mu_1(A) = \mu_1(\{x : f(x) \in A\}), \forall A \in \sigma_2.$$

Let $\gamma \in \mathcal{W}_2$ be an atomless measure, that is, it possesses a continuous cumulative distribution function $F_\gamma$, then the tangent space at $\gamma$ is defined as follows.

**Definition 2.1.**

$$\mathrm{Tan}_\gamma = \overline{\{t(T_\gamma^\mu - id) : \mu \in \mathcal{W}_2,\ t > 0\}}^{\mathcal{L}_\gamma^2},$$

*where* $T_\gamma^\mu = F_\mu^{-1} \circ F_\gamma$.

Note that $\mathrm{Tan}_\gamma$ is endowed with the inner product $\langle \cdot, \cdot \rangle_\gamma$ defined by

$$\langle f, g \rangle_\gamma := \int_{\mathbb{R}} f(x)g(x)\, d\gamma(x),\ f, g \in \mathcal{L}_\gamma^2(\mathbb{R}),$$

and the induced norm $\| \cdot \|_\gamma$. By the definition of pushforward, it is easy to verify that $T_\gamma^\mu$ pushforwards $\mu$ to $\gamma$. In fact, it is the optimal pushforward from $\mu$ to $\gamma$ in the sense that it results in the least transport cost. The technical explanation see for example (Panaretos and Zemel, 2020, Chapter 1).

We now define the exponential and logarithmic maps, which are the ways to communicate between a Tangent space and its Wasserstein space.

**Definition 2.2.** *The logarithmic map* $\mathrm{Log}_\gamma : \mathcal{W}_2 \to \mathrm{Tan}_\gamma$ *is defined as*

$$\mathrm{Log}_\gamma \mu = T_\gamma^\mu - id.$$

*The exponential map* $\mathrm{Exp}_\gamma : \mathrm{Tan}_\gamma \to \mathcal{W}_2$ *is defined as*

$$\mathrm{Exp}_\gamma g = (g + id)\#\gamma.$$

We can see that, given a reference measure, all the measures in the nonlinear Wasserstein space can be mapped into the linear Tangent space of the reference. This implies that we can construct the extended AR model in a Tangent space. Before we move onto the proposed model in the next section, lastly, we introduce the notion of Fréchet mean, which will replace the classical expectation.

**Definition 2.3.** *Let* $\mu_1, \ldots, \mu_T$ *be measures in* $\mathcal{W}_2$. *The empirical Fréchet mean of* $\mu_1, \ldots, \mu_T$, *denoted by* $\bar{\mu}$, *is defined as the unique minimizer of*

$$\min_{\nu \in \mathcal{W}_2} \frac{1}{T} \sum_{t=1}^{T} d_W^2(\mu_t, \nu).$$

**Definition 2.4.** *A random measure* $\boldsymbol{\mu}$ *is any measurable map from a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *to the metric space* $\mathcal{W}_2$, *endowed with its Borel* $\sigma$*-algebra.*

**Definition 2.5.** *Let* $\boldsymbol{\mu}$ *be a random measure from probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *to* $\mathcal{W}_2$. *Assume that* $\boldsymbol{\mu}$ *is square integrable, namely* $\mathbb{E}d_W^2(\boldsymbol{\mu}, \nu) < \infty$ *for some (thus for all)* $\nu \in \mathcal{W}_2$. *Then, the population Fréchet mean of* $\boldsymbol{\mu}$, *denoted by* $\mu_\oplus$, *is defined as the unique minimizer of*

$$\min_{\nu \in \mathcal{W}_2} \mathbb{E}\left[d_W^2(\boldsymbol{\mu}, \nu)\right].$$

For our Wasserstein distance, $\bar{\mu}$ and $\mu_\oplus$ admit simple expressions through their quantile functions:

$$F_{\bar{\mu}}^{-1}(p) = \frac{1}{T} \sum_{t=1}^{T} F_{\mu_t}^{-1}(p),\ F_{\mu_\oplus}^{-1}(p) = \mathbb{E}\left[F_{\boldsymbol{\mu}}^{-1}(p)\right],\ p \in (0, 1). \tag{2.3}$$

## 3 Wasserstein multivariate auto-regressive Models

Given the data of a multivariate distributional time series, $\boldsymbol{\mu}_t^i$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$, we primarily assume that

**Assumption A1.** $\boldsymbol{\mu}_t^i \in \mathcal{W}_2(\mathbb{R})$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$, *and* $\mathbb{E}_\oplus \boldsymbol{\mu}_t^i = \mu_{i,\oplus}, t \in \mathbb{Z}$.

We now extend the VAR(1) models for such data in two steps. In the first step, we will fix a reference measure, on which we will set up a Tangent space. A common choice in literature is Fréchet mean, such as the population Fréchet mean for a set of iid samples. Similarly, the previous works on univariate distributional time series took the population Fréchet mean of the series $\mathbb{E}_\oplus \boldsymbol{\mu}_t$ as the reference. However, as we deal with $N$ series, we have $N$ population Fréchet means, which makes the natural choice not obvious. Thus in this step, we are inspired by the centering operation $\boldsymbol{x}_{j,t-1} - u_j$ in the VAR(1) models and propose a way to center our samples $\boldsymbol{\mu}_t^i$, so that the centered samples will share the common Fréchet mean. The proposed centering applies to general random distributions. Given a random distribution $\boldsymbol{\mu}$ with its quantile function $\boldsymbol{F}^{-1}$, the centered distribution $\tilde{\boldsymbol{\mu}}$ is defined by its quantile function $\tilde{\boldsymbol{F}}^{-1}$ as

$$\tilde{\boldsymbol{F}}^{-1} = \boldsymbol{F}^{-1} \circ (F_\oplus^{-1})^{-1}, \tag{3.1}$$

where $F_\oplus^{-1}(p) = \mathbb{E}\boldsymbol{F}^{-1}(p)$, according to the fact given in Equation (2.3), $F_\oplus^{-1}$ is actually the quantile function of Fréchet mean $\mathbb{E}_\oplus \boldsymbol{\mu}$. Our goal is to make the centered distribution have $U(0,1)$ as its Fréchet mean, that is $\mathbb{E}_\oplus \tilde{\boldsymbol{\mu}} = U(0,1)$. To this end, we need to assume[3] that $\boldsymbol{\mu}$ is supported on $[0,1]$, essentially a closed bounded interval of $\mathbb{R}$. We need to apply this centering transformation on each distribution in our data, thus we add the following data assumption.

**Assumption A2.** *All* $\boldsymbol{\mu}_t^i$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$ *are supported on* $[0,1]$.

Therefore, the centering operation in our extended VAR(1) models is given by

$$\tilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{F}_{i,t}^{-1} \circ (F_{i,\oplus}^{-1})^{-1}, \tag{3.2}$$

where $\boldsymbol{F}_{i,t}^{-1}$ and $\boldsymbol{\mu}_t^i$ are respectively the quantile functions of $\boldsymbol{\mu}_t^i$ and $\mu_{i,\oplus}$, and $\tilde{\boldsymbol{F}}_{i,t}^{-1}$ defines the centered distribution $\tilde{\boldsymbol{\mu}}_t^i$. Then in the second step, we will set up the regressive formula for $\tilde{\boldsymbol{\mu}}_t^i$. The common Fréchet mean, namely $U(0,1)$, therefore will be taken as the reference measure. We propose an extension of VAR(1) model as

$$\tilde{\boldsymbol{\mu}}_t^i = \boldsymbol{\epsilon}_{i,t} \# \operatorname{Exp}_{Leb} \left( \sum_{j=1}^N A_{ij} \operatorname{Log}_{Leb} \tilde{\boldsymbol{\mu}}_{t-1}^j \right), \quad t \in \mathbb{Z},\ i = 1, \ldots, N,$$

where $\{\boldsymbol{\epsilon}_{i,t}\}_{i,t}$ are i.i.d. random distortion functions taking values in the space of extended quantile functions

$$\Pi = \{F^{-1} : [0,1] \to [0,1], \text{ such that } F^{-1}\big|_{(0,1)} \in \operatorname{Log}_{Leb} \mathcal{W} + id,$$
$$F^{-1}(0) := \inf\{x \in [0,1] : F(x) > 0\}, \text{and} F^{-1}(1) := \sup\{x \in [0,1] : F(x) < 1\}\},$$

endowed with $\|\cdot\|_{Leb}$ and the induced Borel algebra, $\boldsymbol{\epsilon}_{i,t}$ is almost surely independent of $\boldsymbol{\mu}_{t-1}^i$, $i = 1, \ldots, N$, for all $t \in \mathbb{Z}$, and

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{i,t}(x)\right] = x,\ x \in [0,1].$$

Note that, we first map the predictor measures $\tilde{\boldsymbol{\mu}}_{t-1}^j$ into the Tangent space of $U(0,1)$ namely the Lebesgue measure over $[0,1]$ as in our notation. Then over the corresponding Tangent vectors $\operatorname{Log}_{Leb} \tilde{\boldsymbol{\mu}}_{t-1}^j$, we apply the same regressive operation as in classical VAR(1) model. The proposed model is not yet identifiable since the exponential map is not injective. Therefore, we need the following assumption.

**Assumption A3.** $\sum_{j=1}^N A_{ij} \leqslant 1$ *and* $0 \leqslant A_{ij} \leqslant 1$.

[3] For the technical explanation, we refer to Jiang (2022).

Given the identifiability, the model allows another representation in terms of quantile functions as follows.

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} \left( \widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right], \quad t \in \mathbb{Z},\, i = 1, \ldots, N, \tag{3.3}$$

where $\widetilde{\boldsymbol{F}}_{i,t}^{-1}$ is the quantile function of $\widetilde{\boldsymbol{\mu}}_{i,t}^{-1}$.

# 4 Existence, uniqueness and stationarity

Now, we present the theoretical properties of the proposed model. The main tool in the derivation is a general result from Wu and Shao (2004), where conditions for an iterated random function (IRF) system in a metric space to be stable are given. Applying the result allows us to show the existence of the solution of our system. Then we build from that, and we are able to show the uniqueness and stationarity of the solution. We rely on the quantile representation (3.3) in the theoretical development by considering it as an IRF system in the following metric space.

$$(\mathcal{X}\,,\, d) := (\mathcal{T}\,,\, \|\cdot\|_{Leb})^{\otimes N}\,,$$

where $\mathcal{T} := \mathrm{Log}_{Leb}\, \mathcal{W}_2(\mathbb{R}) + id$ is the space of all quantile functions of $\mathcal{W}_2(\mathbb{R})$, equipped with the norm $\|\cdot\|_{Leb}$. Thus, we have

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{N} \|\mathbf{X}_i - \mathbf{Y}_i\|_{Leb}^2}, \quad \mathbf{X} = (\mathbf{X}_i)_{i=1}^{N} \in \mathcal{X},\ \mathbf{Y} = (\mathbf{Y}_i)_{i=1}^{N} \in \mathcal{X}.$$

We propose the following assumptions, which allow our system to satisfy the stability conditions required by Wu and Shao (2004).

**Assumption A4.** $\mathbb{E}\left[\boldsymbol{\epsilon}_{i,t}(x) - \boldsymbol{\epsilon}_{i,t}(y)\right]^2 \leqslant L^2(x-y)^2,\ \forall x, y \in [0,1], \quad t \in \mathbb{Z},\, i = 1, \ldots, N,$

**Assumption A5.** $\|A\|_2 < \frac{1}{L}$.

Note that, Assumption A4 implies that $\boldsymbol{\epsilon}_{i,t}$ is $L$-Lipschitz in expectation. For increasing functions from $[0,1]$ to $[0,1]$, the smallest $L$ is 1 that is attained by the identity function. Therefore, Assumption A5 implies that $\|A\|_2 < 1$, which is the stability condition for VAR(1) models. Theorem 4.1 states the existence and uniqueness results.

**Theorem 4.1.** *Under Assumptions A3, A4 and A5, the IRF system (3.3) almost surely admits a solution $\mathbf{X}_t$, $t \in \mathbb{Z}$, with the same marginal distribution $\pi$, namely, $\mathbf{X}_t \overset{d}{=} \pi$, $\forall\, t \in \mathbb{Z}$, where the notation $\overset{d}{=}$ means equality in distribution. Moreover, if there exists another solution $\boldsymbol{S}_t$, $t \in \mathbb{Z}$, then for all $t \in \mathbb{Z}$*

$$\mathbf{X}_t \overset{d}{=} \boldsymbol{S}_t, \quad \textit{almost surely.}$$

We are able to show that this unique (in the sense of distributions) solution is furthermore stationary[4] as a functional process in a Hilbert space. To this end, we need to assume that there is an underlying Hilbert space associated to $(\mathcal{X}\,,\, d)$, with its inner product inducing $d$ as the norm. Such Hilbert space exists, with corresponding inner product given by

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{N} \langle \mathbf{X}_i, \mathbf{Y}_i \rangle_{Leb}\,.$$

We recall the conventional definition of stationarity for process in a separable Hilbert space, see for example Zhang et al. (2021, Definition 2.2).

[4]Different from stability, stationarity requires the second order invariance with respect to some inner product, thus it is a notion defined in a Hilbert space.

**Definition 4.1.** *A random process* $\{\mathbf{V}_t\}_t$ *in a separable Hilbert space* $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ *is said to be stationary if the following properties are satisfied.*

1. $\mathbb{E}\left\|\mathbf{V}_t\right\|^2 < \infty$.
2. *The Hilbert mean* $U := \mathbb{E}\left[\mathbf{V}_t\right]$ *does not depend on* $t$.
3. *The auto-covariance operators defined as*

$$\mathcal{G}_{t,t-h}(V) := \mathbb{E}\langle \mathbf{V}_t - U, V\rangle (\mathbf{V}_{t-h} - U), \quad V \in \mathcal{H},$$

*do not depend on* $t$, *that is* $\mathcal{G}_{t,t-h}(V) = \mathcal{G}_{0,-h}(V)$ *for all* $t$.

Then, Theorem 4.2 below gives the stationarity result.

**Theorem 4.2.** *The unique solution given in Theorem 4.1 is stationary as a random process in* $(\mathcal{X}, \langle\cdot,\cdot\rangle)$ *in the sense of Definition 4.1.*

# 5 Estimation of the regression coefficients

In this section, we develop a consistent estimator of coefficient $A$, given $T+1$ samples $\boldsymbol{\mu}_t^i$, $t = 0, 1, \ldots, T, i = 1, \ldots, N$. We calculate the estimator in two steps as well. Firstly, we estimate the exact centered data $\tilde{\boldsymbol{\mu}}_t^i$ defined in Equation (3.2) using empirical Fréchet means. We denote the estimates by $\widehat{\boldsymbol{\mu}}_t^i$, which are defined by their quantile functions as:

$$\widehat{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{F}_{i,t}^{-1} \circ (\boldsymbol{F}_{\bar{\mu}_i}^{-1})^{-1}, \quad \text{where } \boldsymbol{F}_{\bar{\mu}_i}^{-1} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{F}_{\mu_{i,t}}^{-1},$$

Then the proposed estimator is defined through the following least squares formula.

$$\widehat{\boldsymbol{A}}_{i:} = \underset{A_{i:} \in B_+^1}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{\boldsymbol{F}}_{i,t}^{-1} - \sum_{j=1}^{N} A_{ij}\left(\widehat{\boldsymbol{F}}_{j,t-1}^{-1} - id\right) - id \right\|_{Leb}^2, \quad i = 1, \ldots, N, \tag{5.1}$$

where $B_+^1$ is $N$-dimensional simplex, that is the nonnegative orthant of the $\ell_1$ unit ball $B^1$ in $\mathbb{R}^N$, corresponding to Assumption A3. Thus, an important advantage of this constraint is to promote sparsity in $\widehat{\boldsymbol{A}}_{i:}$. When using the proposed model in graph learning, the retrieved (directed) graph from $\widehat{\boldsymbol{A}}$ will be naturally sparse. The optimisation problem (5.1) can be solved by the accelerated projected gradient descent (Parikh and Boyd, 2014, Chapter 4.3). The projection onto $B_+^1$ is given in Thai et al. (2015). Theorem 5.1 shows that the proposed estimator is consistent.

**Theorem 5.1.** *Assume that* $\boldsymbol{\mu}_t^i$, $i = 1, \ldots, N$ *satisfy Assumption A1 for* $t = 0, 1, \ldots, T$, *and the transformed sequence* $\widetilde{\boldsymbol{F}}_{i,t}^{-1}$ *satisfies Model (3.3) with Assumption A3 true. Suppose additionally that* $(\widetilde{\boldsymbol{F}}_{i,0}^{-1})_{i=1}^N \overset{d}{=} \pi$ *with* $\pi$ *the stationary distribution defined in Theorem 4.1. Given Assumptions A4 and A5 hold true, and the following* $N \times N$ *matrix* $\Gamma(0)$ *is nonsingular*

$$[\Gamma(0)]_{j,l} = \mathbb{E}\langle \widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id\,,\, \widetilde{\boldsymbol{F}}_{l,t-1}^{-1} - id\rangle_{Leb},$$

*we have*

$$\widehat{\boldsymbol{A}} - A \xrightarrow{p} 0.$$

## 6 Numerical experiment

In this section, we fit our model with a real data set which records annual age distributions of countries/states. The data set has been illustrated in Figure 1 in the introduction. We represent the distribution of age population, of country $i$, at year $t$, by $\boldsymbol{\mu}_t^i$, with $T = 1, ..., 40$ and $i = 1, ..., 34$. To justify the proposed model and to illustrate its application in graph learning from distributional time series, we visualize the estimation of regression coefficients $A$ on the real geographical map in Figure 2, so as to inspect the learned patterns. For details of the data and experimental settings as well as more experiment results, we refer to Jiang (2022).

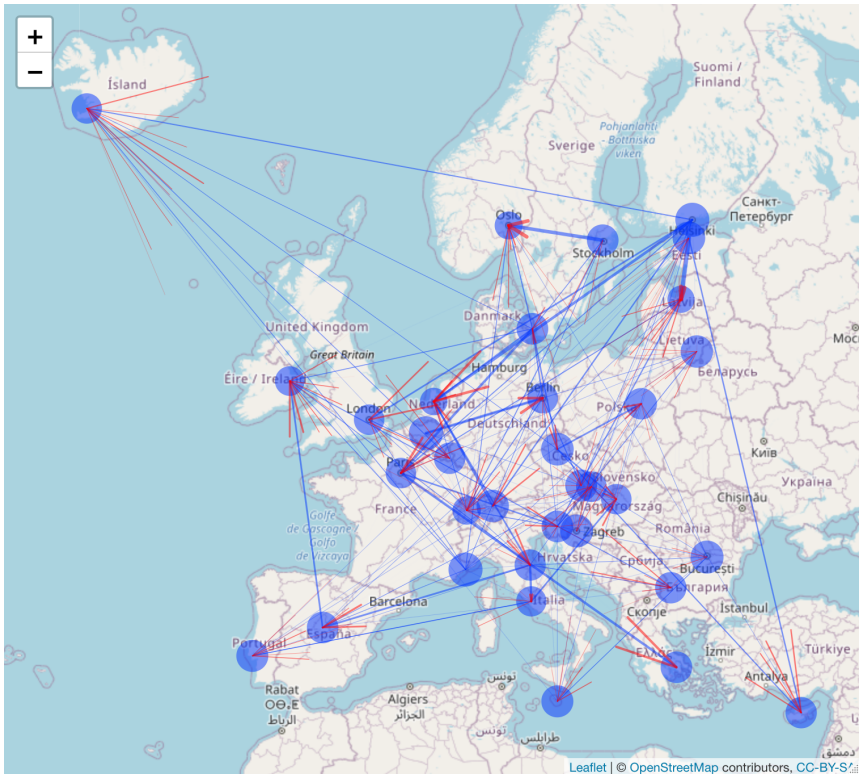



Figure 2: *Inferred age structure graph.* The non-zero coefficients $A_{ij}$ are represented by the weighted directed edges from node $j$ to node $i$. Thicker arrow corresponds to larger weights. The blue circles around nodes represent the weights of self-loop.

We can notice that for all countries $i \in \{1, ..., 34\}$, the coefficients of self-loop $A_{ii}$ dominate others $A_{ij}$, $i \neq j$. This is because the age structure of a country does not change much from one year to another. On the other hand, this also implies the age structure differs largely across countries. Nevertheless, there are still significant links between different countries' age distributions. The first two largest international coefficients are: Estonia → Latvia, and Sweden → Norway. To justify these inferred patterns, we plot the time series of the four countries in Figure 3. We can see that the countries on the same row (their regression coefficients have large values) have similar time series along time, by contrast, those on the same column (their regression coefficients have small or zero values) differ a lot. All these observations strongly support the usefulness of our model.

## References

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures.* Springer Science & Business Media, 2008.

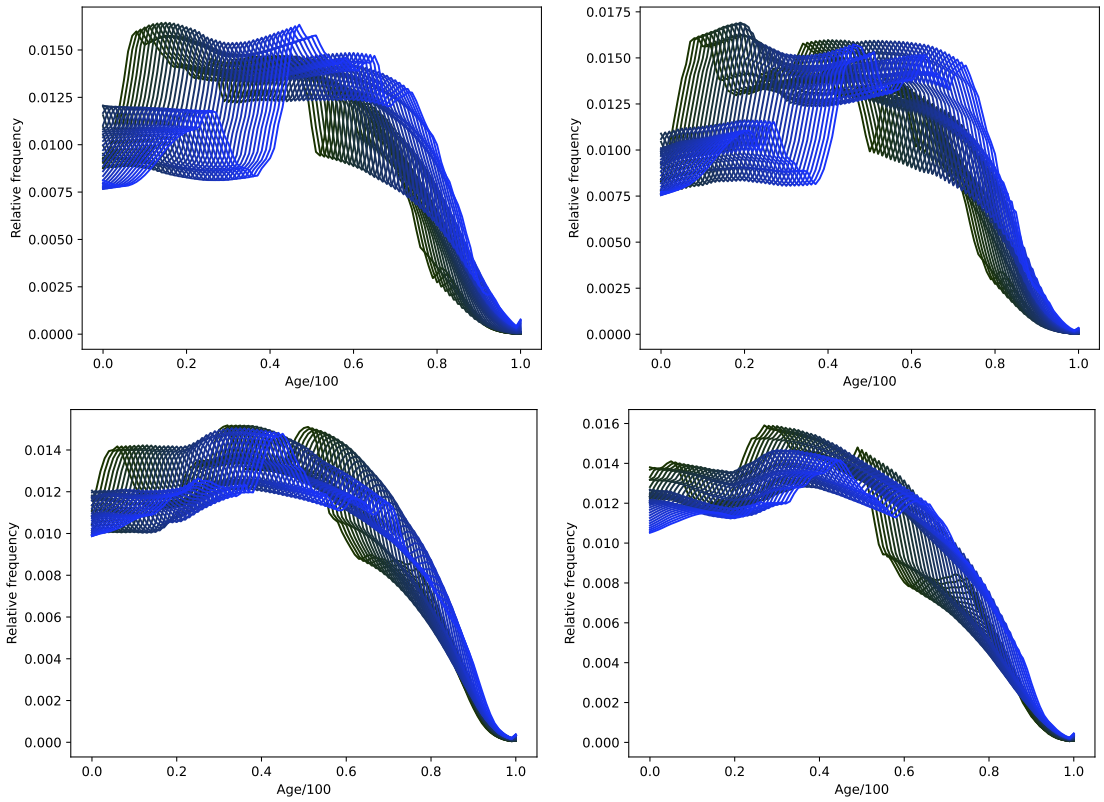



Figure 3: *Evolution of age structure from* 1996 *to* 2036 *(projected) of Estonia (left upper), Latvia (right upper), Sweden (left bottom) versus Norway (right bottom).* Each curve connects the 101 relative frequencies from $0, 1/100, 2/100, ..., 1$, which represents the age structure of a considered year. Lighter curves correspond to more recent years.

J. Bigot. Statistical data analysis in the wasserstein space. *ESAIM: ProcS*, 68:1–19, 2020.

J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1–26, 2017.

D. Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, 2000.

Y. Chen, Z. Lin, and H.-G. Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

Y. Jiang. Wasserstein multivariate auto-regressive models for modeling distributional time series and its application in graph learning. *arXiv preprint arXiv:2207.05442*, 2022.

P. Kokoszka, H. Miao, A. Petersen, and H. L. Shang. Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, 35(4):1304–1317, 2019.

S. Mazzuco and B. Scarpa. Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):187–203, 2015.

V. M. Panaretos and Y. Zemel. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

V. M. Panaretos and Y. Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

A. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

A. Petersen and H.-G. Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2): 339–351, 2019.

A. Petersen, C. Zhang, and P. Kokoszka. Modeling Probability Density Functions as Data Objects. *Econometrics and Statistics*, 21(C):159–178, 2022.

H. L. Shang and S. Haberman. Forecasting age distribution of death counts: An application to annuity pricing. *Annals of Actuarial Science*, 14(1):150–169, 2020.

J. Thai, C. Wu, A. Pozdnukhov, and A. Bayen. Projected sub-gradient with $\ell_1$ or simplex constraints via isotonic regression. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 2031–2036. IEEE, 2015.

W. B. Wu and X. Shao. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2): 425–436, 2004.

Y. Zemel and V. M. Panaretos. Fréchet means and procrustes analysis in wasserstein space. 2019.

C. Zhang, P. Kokoszka, and A. Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 2021.

C. Zhu and H.-G. Müller. Autoregressive optimal transport models. *arXiv preprint arXiv:2105.05439*, 2021.

# Interprétation des modèles de régression compositionnelle basées sur les ratio de paires de composantes

Lukas Dargel [1] & Christine Thomas-Agnan [2]

[1] *Toulouse School of Economics, France, christine.thomas@tse-fr.eu*
[2] *Toulouse School of Economics, France, lukas.dargel@tse-fr.eu*

**Résumé.** L'interprétation des modèles de régression comportant des vecteurs de parts aussi nommés compositions en tant que variable réponse et/où explicative a été abordée sous divers angles. Les premières approches de la littérature se font dans l'espace dit des coordonnées c'éest-à-dire après transformation des variables de composition par des transformations de type log-ratio. Etant donné que ces modèles sont non linéaires par rapport aux opérations classiques de l'espace $\mathbb{R}^D$, une autre approche a été proposée basée sur des incréments infinitésimaux et sur des dérivées au sens du simplexe. Cette dernière conduit à une interprétation à base d'élasticités ou de semi-élasticités. elles se font dans l'espace d'origine du vecteur de composition et sont indépendantes de toute transformation. Après un rappel sur ces deux points de vue, nous montrons que certaines fonctions des élasticités ou semi-élasticités sont constantes au travers de l'échantillon, ce qui en fait des paramètres naturels pour l'interprétation des modèles de régression compositionnelle. Ces paramètres sont liés à des variations relatives de ratio de deux composantes des variables réponse et/où explicatives. Nous proposons également des approximations de ces quantités pour un petit increment et montrons leur lien avec les paramètres naturels précédemment cités ce qui conduit à des interprétations libres de toute transformation et portant sur les variations exprimées dans l'espace d'origine. Nous utilisons un jeu de données sur les élections présidentielles françaises de 2022 pour illustrer chaque type d'interprétation.

**Mots-clés.** modèles de régression compositionnelle, ratios de paires, mesures d'impact, dérivées dans le simplexe

**Abstract.** The interpretation of regression models with compositional vectors as response and/or explanatory variables has been approached from different perspectives. The first approaches that appear in the literature are done in coordinate space after some log-ratio transformation of the compositional vectors. Considering the fact that these models are non-linear with respect to classical operations of the real space, another approach has been proposed based on infinitesimal increments or derivatives understood in a simplex sense, leading to elasticities or semi-elasticities interpretations in the original share space that have the advantage of being independent of any log-ratio transformations. After briefly reviewing these two points of view, we show that some functions of elasticities or semi-elasticities are constant throughout the sample observations, which makes them natural parameters for interpreting CoDa models. These parameters are linked to relative variations of pairwise share ratios of the response and/or of the explanatory variables. We derive approximations of share ratio variations and link them to these natural parameters leading to transformation-free

interpretations in the original share space. We use a real dataset on the French presidential election to illustrate each type of interpretation in detail.

**Keywords.** Compositional data regression, pairwise share ratio, impact measures, simplicial derivative.

# 1 Difficultés de l'interprétation d'un modèle de régression comportant des vecteurs de parts

L'interprétation des paramètres dans un modèle de régression est une étape essentielle pour comprendre l'impact marginal des changements d' une variable explicative sur la variable de réponse. Rappelons tout d'abord que, dans un modèle de régression linéaire classique expliquant la réponse Y par un ensemble de variables explicatives X, l'espérance conditionnelle $\mathbb{E}(Y \mid X)$ est une fonction linéaire de $X$. Par conséquent, pour toute variable explicative spécifique $X_k$, nous pouvons comprendre le paramètre $\beta_{X_k}$ de $X_k$ comme l'accroissement additif de $\mathbb{E}(Y \mid X_k)$ lorsque $X_k$ augmente d'une unité (accroissement fini), toutes les autres variables explicatives restant fixées (ceteris paribus), ou alternativement comme la dérivée de $\mathbb{E}(Y \mid X_k)$ par rapport à $X_k$ (accroissement infinitésimal). En économétrie, l'interprétation basée sur la dérivée est connue sous le nom d'effet marginal et les deux points de vue (accroissements finis et infinitésimaux) coïncident pour les modèles linéaires. Cependant, les modèles de régression CoDa impliquent des variables vecteurs de parts, également appelées variables compositionnelles du côté droit et/ou du côté gauche de l'équation de régression, ce qui implique qu'au moins certains paramètres ou variables du modèle sont à valeur dans un simplexe. Pour cette raison, les modèles CoDa ne sont pas linéaires pour la structure de l'espace vectoriel de l'espace réel, et c'est pourquoi les premières interprétations dans la littérature sont effectuées dans l'espace dit des coordonnées après une certaine transformation en log-ratio des vecteurs de parts. L'absence de linéarité du côté gauche (lorsque la réponse est compositionnelle) peut être résolue en adaptant la définition de l'espérance aux variables compositionnelles (voir par exemple [Pawlowsky-Glahn et al., 2015a]) comme cela a déjà été mentionné dans [Morais et Thomas-Agnan, 2021]. Comme nous le montrons dans ce texte, l'absence de linéarité du côté droit (lorsque l'explicative est compositionnelle) se traduit par le fait qu'on ne peut pas changer une composante d'un vecteur de parts en maintenant les autres composantes constantes: cette difficulté peut être résolue en considérant des accroissements linéaires des variables explicatives où la linéarité est comprise par rapport à la géométrie du simplexe introduite par Aitchison et définie par les opérations de perturbation et d'exponentiation.

## 2 Interprétations classiques et nouvelles d'un modèle de régression CoDa

Ce travail présente et illustre des interprétations classiques mais aussi nouvelles de l'impact des variables explicatives dans les modèles de régression comportant des vecteurs de parts. Les exemples de l'exposé seront centrés sur l'interprétation d'une régression expliquant les parts de vote pour différentes candidats ou groupes de candidats ainsi que le taux de participation lors du premier tour de l' élection présidentielle de 2022 en fonction de variables socio-économiques. L'utilisation de techniques de données compositionnelles pour analyser les données électorales est naturelle et plusieurs références peuvent être trouvées comme [Katz and King, 1999] et [Nguyen et al., 2022]. Après avoir rappelé les bases de la régression CoDa, notre premier objectif est de présenter une illustration complète des interprétations des élasticités/semi-élasticités de [Morais et al., 2018] et [Morais et Thomas-Agnan, 2021]. Nous discutons dans un deuxième temps des interprétations s'appuyant sur l'espace des coordonnées et de montrons l'avantage de l'interprétation dans le simplexe. Enfin, notre dernier objectif est de proposer une nouvelle interprétation basée sur les variations des ratios de paires de composantes dans l'espace du simplexe.

Les résultats mathématiques utilisés pour obtenir ces interprétations sont des approximations de parts et de ratios de parts dues à de petits accroissements linéaires d'un vecteur de parts explicatif le long d'un chemin linéaire dans le simplexe. Comprendre les chemins linéaires dans le simplexe est utile pour manipuler les variations linéaires des variables explicatives compositionnelles ainsi que pour interpréter les variations des vecteurs de parts de la variable réponse lorsque celle-ci est compositionnelle. Nous établissons une nouvelle formule de Taylor pour approximer une fonction d'un simplexe vers un autre simplexe le long d'un chemin linéaire dans une direction générale du simplexe.

Dans le cas de la réponse scalaire, nous montrons que toutes les approches sont équivalentes mais la nôtre permet de considérer des accroissements plus généraux des variables explicatives compositionnelles. Dans le cas de la variable réponse compositionelle, nous montrons que les variations infinitésimales des ratio de paires de parts sont indépendantes de l'individu dans l'échantillon, en faisant ainsi des paramètres essentiels pour l'interprétation de ces modèles. Par ailleurs ces interprétations sont plus faciles à comprendre pour les utilisateurs.

Tous les outils présentés sont disponibles dans le package R **CoDaImpact**, qui peut être utilisé en coordination avec le package R **compositions** [van den Boogaart et al.,2023]. Les illustrations sont disponibles dans la vignette

`\url{https://github.com/LukeCe/CoDaImpact}.`

## Bibliographie

Aitchison, J., et Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. Biometrika, 71(2), 323-330.

Coenders, G., et Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT-Statistics and Operations Research Transactions, 201-220.

Dargel, L., et Thomas-Agnan, C. (2023). Pairwise share-ratio interpretations of compositional regression models. TSE working paper n. 23-1456, Juillet 2023, révisé le 20 septembre 2023.

Morais, J., et Thomas-Agnan, C. (2021). Impact of covariates in compositional models and simplicial derivatives. Austrian Journal of Statistics, 50(2), 1-15.

Muller I. et al. (2018). Interpretation of Compositional Regression with Application to Time Budget Analysis. Austrian Journal of Statistics, (2), 3-19.

van den Boogaart K.et al. (2021). Classical and robust regression analysis with compositional data. Mathematical geosciences, 53:823–858.

van den Boogaart, K., Tolosana-Delgado, R., et Bren, M. (2023). compositions: Compositional Data Analysis. R package version 2.0-5

# Spatial Autoregressive Model on a Dirichlet distribution

Teo Nguyen[1,2] & Sarat Moka[3] & Kerrie Mengersen[1,4] & Benoit Liquet[1,2]

[1] *Université de Pau et des Pays de l'Adour, Anglet, France*
[2] *Macquarie University, Sydney, Australia*
[3] *University of New South Wales, Sydney, Australia*
[4] *Queensland University of Technology, Brisbane, Australia*

teo.nguyen@univ-pau.fr
s.moka@unsw.edu.au
k.mengersen@qut.edu.au
benoit.liquet@univ-pau.fr

**Résumé.** Les données de composition sont largement utilisées dans divers domaines tels que l'écologie, la géologie, l'économie et la santé publique, car elles représentent les proportions ou les pourcentages des différents éléments composant un ensemble. Toutefois, en raison de leur nature relative et de leur contrainte de vivre dans un simplexe, les méthodes statistiques traditionnelles ne sont pas directement applicables aux données de composition (Aitchison 1982). Des dépendances spatiales existent souvent dans les données de composition, en particulier lorsque les composantes représentent des différentes répartitions des terres ou bien des variables écologiques. L'auto-corrélation spatiale peut résulter de conditions environnementales communes ou de la proximité géographique. Il est donc essentiel d'incorporer des informations spatiales dans l'analyse statistique des données de composition afin d'obtenir des résultats précis et fiables. Pour traiter les données compositionnelles, la distribution de Dirichlet est couramment utilisée car son support est un vecteur compositionnel. Maier (2014) a proposé un modèle de régression pour les données à distribution de Dirichlet, mais ce modèle ne prend pas en compte les dépendances spatiales, ce qui limite son applicabilité aux problèmes spatiaux. Dans cette étude, nous présentons un modèle autorégressif spatial pour des données suivant une distribution de Dirichlet qui incorpore des dépendances spatiales entre les observations. Nous développons un estimateur du maximum de vraisemblance sur une fonction de densité de Dirichlet qui inclut un terme dit de "spatial lag". Nous comparons ce modèle autorégressif spatial au même modèle sans "spatial lag" et testons les deux modèles sur des ensembles de données synthétiques et réelles. Différentes matrices de poids spatiales sont utilisées pour tenir compte de leur effet sur l'ensemble des données synthétiques. Les résultats démontrent que l'incorporation des dépendances spatiales peut améliorer la performance du modèle et confirment que l'efficacité dépend de la définition de la matrice des poids (Anselin 1988). En tenant compte des relations spatiales entre les observations, notre modèle fournit des résultats plus précis et plus fiables pour l'analyse des données de composition. Les recherches futures pourraient explorer plus en détail l'application du modèle proposé dans différents domaines et étudier d'autres matrices de poids pour l'analyse des données de composition dans divers contextes spatiaux.

**Mots-clés.** données de composition, modèle autoregressif spatial, régression de dirichlet.

**Abstract.** Compositional data are widely utilized in various fields, such as ecology, geology, economics, and public health, as they effectively represent proportions or percentages of different components in a whole. However, due to their relative nature and the constraint of lying on a simplex, traditional statistical methods are not directly applicable to compositional data (Aitchison 1982). Spatial dependencies often exist in compositional data, particularly when the components represent different land uses or ecological variables. Spatial autocorrelation can arise from shared environmental conditions or geographical proximity. Therefore, it is essential to incorporate spatial information into the statistical analysis of compositional data to obtain accurate and reliable results. To handle compositional data, the Dirichlet distribution is commonly used as its support is a compositional vector. Maier (2014) proposed a regression model for Dirichlet-distributed data, but this model does not consider spatial dependencies, which limits its applicability in spatial problems. In this study, we introduce a spatial autoregressive model for Dirichlet-distributed data that incorporates spatial dependencies between observations. We develop a maximum likelihood estimator on a Dirichlet density function that includes a spatial lag term. We compare this spatial autoregressive model with the same model without spatial lag and test both models on synthetic and real datasets. Different spatial weights matrices are employed to account for their effect on the synthetic dataset. The results demonstrate that incorporating spatial dependencies can improve the performance of the model and confirm that the efficiency depends on the definition of the spatial weights matrix (Anselin 1988). By considering the spatial relationships among observations, our model provides more accurate and reliable results for the analysis of compositional data. Future research could further explore the application of the proposed model in different fields and investigate alternative spatial weights matrices for compositional data analysis in diverse spatial contexts.

**Keywords.** compositional data, dirichlet regression, spatial autoregressive model.

# 1 Introduction

Compositional data, widely used in different fields such as ecology, geology or economics, are data able to represent proportions or percentages of different components in a whole. We define a $D$-part compositional vector as a vector $y = (y_1, y_2, \ldots, y_D) \in \mathbb{R}^D$ such that,

$$\begin{cases} y_i \geq 0, \quad \forall\, i \in \{1, 2, \ldots, D\}, \\ \sum_{i=1}^{D} y_i = 1. \end{cases}$$

Compositional vectors lie on a simplex $S^D$, where traditional statistical methods cannot be applied directly (Aitchison 1982).

One of the most commonly used probability distributions for compositional data is the Dirichlet distribution, as a Dirichlet distribution of parameter $\alpha \in \mathbb{R}^D$ will generate a $D$-part

compositional vector. Maier (2014) proposed a regression model for Dirichlet-distributed data, but this model does not take spatial dependencies into account.

Over the past decades, spatial autoregressive (SAR) models have emerged as powerful tools for analyzing spatially correlated data in various fields, including economics, ecology, and epidemiology. The fundamental idea behind SAR models is that the value of a variable at a particular location is influenced not only by its own characteristics but also by the characteristics of neighboring locations. These models explicitly account for the spatial interdependencies among the observed variables, allowing for a more comprehensive understanding of the underlying spatial processes. While spatial dependencies are often present in compositional data, particularly when the components represent different land uses or ecological variables, only a few studies have developed a SAR model for such data. In these studies, the authors employed either a Bayesian estimation approach to estimate the parameters of a spatial multinomial logit model (Krisztin et al. 2022) or transform the data into the Euclidian space before applying a multivariate regression model (Nguyen et al. 2021), a Gaussian Markov random field (Pirzamanbein et al. 2018) or a multivariate conditionally autoregressive model (Leininger et al. 2013).

Here, we present a spatial autoregressive model for Dirichlet-distributed data. We develop a maximum likelihood estimator that handles the spatial interdependency and demonstrate the effectiveness of our model on one synthetic dataset and two real-world datasets.

# 2 Materials and Methods

We consider the case where the labels of the dataset are compositional. Let $K$ be the number of features, $J$ the number of classes, $n$ the sample size of the dataset. We denote the features of a sample $i$ as $x_i \in \mathbb{R}^K$ and its label $y_i \in S^J$ (i.e., $y_i$ is a compositional vector of dimension $J$). The features (resp., labels) of the whole dataset are then denoted by $X \in \mathbb{R}^{n\times K}$ (resp., $Y \in \mathbb{R}^{n\times J}$). If for a given data row $i$, the label $y_i$ follows a Dirichlet of parameter $\alpha_i \in \mathbb{R}^J$, then the probability density function is

$$f(y_i|\alpha_i) = \frac{\Gamma(\sum_{j=1}^J \alpha_{ij})}{\prod_{j=1}^J \Gamma(\alpha_{ij})} \prod_{j=1}^J y_{ij}^{\alpha_{ij}-1},$$

where $\Gamma$ is the gamma function, $\alpha_i$ is such that $\alpha_{ij} > 0$ for every class $j$. The parameters $\alpha_i$ can be parametrized by $\alpha_i = \phi_i \mu_i$ where $\phi_i \in \mathbb{R}$ is called the precision parameter (or dispersion parameter) and the compositional vector $\mu_i \in S^J$ represents the individual expected values. Hence, the model's predictions $\hat{y}_i$ is given by the estimated values $\hat{\mu}_i$. The parameter $\phi_i$ has an effect on the distinction of the classes. For a fixed $\mu_i$, the smaller $\phi_i$ is, the more likely the point will be distributed around extreme values (the edges of the simplex), while with a high $\phi_i$, the point is more likely to be close to the value of $\mu_i$ (see Figure 1).

All the parameters $\alpha_i$ can be stacked into a matrix $\alpha \in \mathbb{R}^{n\times J}$. Similarly, we can stack all the $\mu_i$ (resp. $\phi_i$) in a matrix $\mu \in \mathbb{R}^{n\times J}$ (resp. a vector $\phi \in \mathbb{R}^n$).

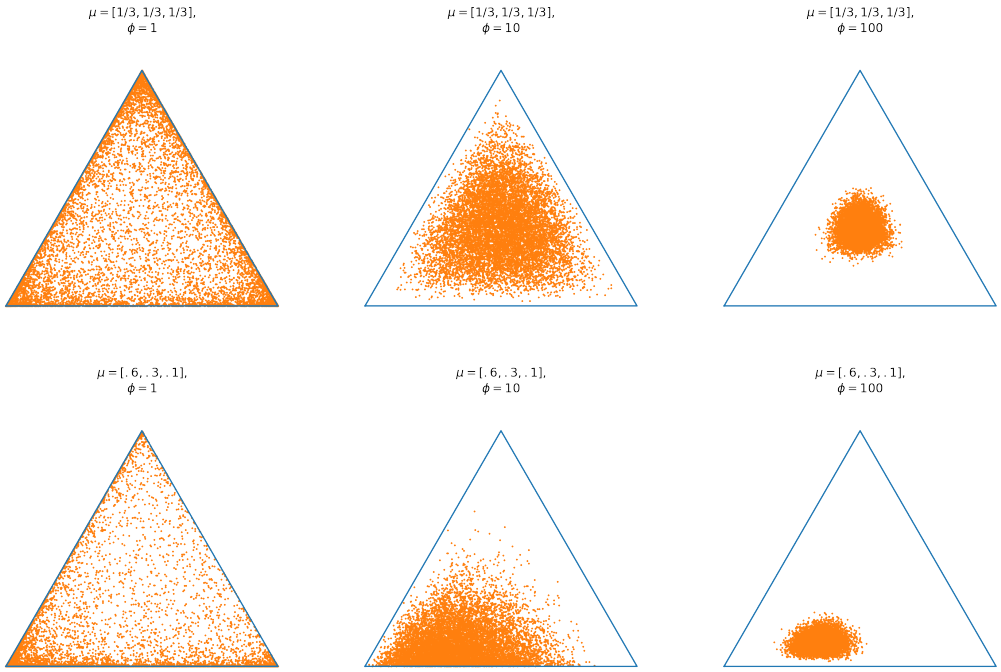



Figure 1: Distribution of 10000 points drawn from a Dirichlet distribution, for different values of $\mu \in S^3$ and $\phi \in \mathbb{R}$.

## 2.1 Maximum likelihood regression without spatial lag

Let $\beta \in \mathbb{R}^{K \times J}$ be a matrix of coefficients. The parameter $\mu \in \mathbb{R}^{n \times J}$ can be defined as depending of $\beta$ and the features $X$ such as

$$\forall i \in [1, \dots, n],\ \forall j \in [1, \dots, J], \quad \mu_{ij} = \frac{\exp(\sum_{k=1}^{K} X_{ik}\beta_{kj})}{\sum_{j'=1}^{J} \exp(\sum_{k=1}^{K} X_{ik}\beta_{kj'})}. \tag{1}$$

Then, let $K_Z \in \mathbb{N}$ where $\mathbb{N}$ denotes the set of nonnegative integers. We introduce a matrix $Z \in \mathbb{R}^{n \times K_Z}$ and a vector $\gamma \in \mathbb{R}^{K_Z}$, that allow to define $\phi \in \mathbb{R}^n$ as

$$\forall i \in [1, \dots, n], \quad \phi_i = \exp([Z\gamma]_i).$$

For any row $i$ and class $j$, we set $\alpha_{ij=}\phi_i\mu_{ij}$, which implies that $\phi_i = \sum_j \alpha_{ij}$. This parametrization is referred to as the *alternative* parametrization in Maier (2014), as opposed to the *common* parametrization where each $\phi_i$ is set to 1.

To ensure the unicity of the solution when maximizing the likelihood, the mapping $\beta \mapsto \mu$ has to be injective, which is ensured by setting a column of $\beta$ as 0, for instance the first column, as done in Maier (2014).

The density function can be rewritten depending on $\mu$ and $\phi$,

$$f(y_i|\mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\prod_{j=1}^{J} \Gamma(\phi_i\mu_{ij})} \prod_{j=1}^{J} y_{ij}^{\phi_i\mu_{ij}-1}. \tag{2}$$

And thus, the log-likelihood of the Dirichlet distribution is,

$$\ell(y|\mu,\phi) = \sum_{i=1}^{n}\left(\ln\Gamma(\phi_i) - \sum_{j=1}^{J}\ln(\Gamma(\phi_i\mu_{ij})) + \sum_{j=1}^{J}((\phi_i\mu_{ij}-1)\ln(y_{ij}))\right) \tag{3}$$

$$= \sum_{i=1}^{n}\left(\ln\Gamma(\phi_i) - \sum_{j=1}^{J}\ln\left(\Gamma\big(\phi_i\frac{\exp([X\beta]_{ij})}{\sum_{j'=1}^{J}\exp([X\beta]_{ij'})}\big)\right) + \sum_{j=1}^{J}\left((\phi_i\frac{\exp([X\beta]_{ij})}{\sum_{j'=1}^{J}\exp([X\beta]_{ij'})}-1\right)\ln(y_{ij}))\right). \tag{4}$$

Because of the $\ln(y_{ij})$ term, $y_{ij}$ needs to be strictly positive. This issue is addressed by using the transformation $y^* = \frac{y(n-1)+1/J}{n}$ (Maier 2014), which ensures that the transformed values are positive and has the property that $\lim_{n\to+\infty} y^* = y$. In the following, we will still denote the data as $y$ and assume it does not contain any zero values, but note that the transformation can be applied if necessary.

Because $\mu$ and $\phi$ are parameterized by $\beta$ and $\gamma$, maximum likelihood estimators $\hat{\beta}$ and $\hat{\gamma}$ are used to estimate these parameters, which then allow us to predict the label of an unseen data point $\tilde{x} \in \mathbb{R}^K$. This prediction is the compositional vector $\tilde{\mu} \in S^J$, computed from (1). The probability vector $\tilde{\mu}$ is considered as being the estimated value of the label.

## 2.2 Maximum likelihood regression with spatial lag

We now introduce a *spatial lag* term through the matrix $M = I_n - \rho W$, where $I_n$ is the identity matrix of size $n$, $\rho \in \mathbb{R}$ is the strength of spatial correlation and $W \in \mathbb{R}^{n\times n}$ is the spatial weights matrix (Anselin 1988). This spatial lag term will allow us to introduce spatial effect in the model. It is common to apply row-normalization on $W$ in order to make its rows sum to 1. Because of this, this matrix is often asymmetric, even though the original non-normalized weights matrix was symmetric.

For a given matrix $\beta$, we redefine $\mu$ as:

$$\forall(i,j), \quad \mu_{ij} = \frac{\exp([M^{-1}X\beta]_{ij})}{\sum_{j'=1}^{J}\exp([M^{-1}X\beta]_{ij'})} = \frac{\exp(\sum_{i'=1}^{n}\sum_{k=1}^{K}M^{-1}_{ii'}X_{i'k}\beta_{kj})}{\sum_{j'=1}^{J}\exp(\sum_{i'=1}^{n}\sum_{k=K}^{n}M^{-1}_{ii'}X_{i'k}\beta_{kj'})}. \tag{5}$$

The introduction of $M$ modifies the computation of the vector $\mu$, and by multiplying $X\beta$ with the inverse of $M$, we introduce the explanatory variables of the neighboring observations.

The value of the spatial correlation parameter $\rho$ needs to be estimated from the data, while $W$ is fixed and has to be defined beforehand. Common choices include distance-based weights or contiguity-based weights (Cliff et Ord 1970). In distance-based weights, the weight between each pair of points is determined by the inverse of the distance between them, while in contiguity-based weights, for each points, the same weight is given to each of its nearest neighbours.

Finally, if we set a matrix $\tilde{X} \in \mathbb{R}^{n\times K}$ such that $\tilde{X} = M^{-1}X$, the loglikelihood remains the same as in (4) provided that we replace the term $X$ with $\tilde{X}$ in this expression.

# 3 Results

We present here the results obtained from applying both the spatial lag model and the non-spatial model to the different synthetic and real datasets. Each dataset is described and some results are presented for each of them.

## 3.1 Synthetic dataset

The synthetic spatially-correlated dataset is generated with 2 features and 3 classes, by varying the number of samples $n$ (50, 200, or 1000) and the values of $\rho$ (0.1, 0.5, or 0.9). The values of $\beta$ and $\gamma$ are predetermined at the beginning of the simulation. In the presented results, we used

$$\beta = \begin{matrix} \\ \text{features} \end{matrix} \overset{\text{classes}}{\begin{bmatrix} 0 & 0 & 0.1 \\ 0 & 1 & -2 \\ 0 & -1 & -2 \end{bmatrix}}, \qquad \gamma = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

These specific values were selected to ensure certain properties in the generated data. The value of $\beta$ has been chosen to ensure that the classes are balanced in $\mu$, i.e., that no class is significantly more frequent or rarer than others. The values of $\gamma$ are chosen in a way that the precision parameter $\phi$ is sufficiently high, so that the distribution of the points is relatively concentrated around their class probabilities, which ensures well-defined class patterns that are more distinguishable.

First, the features matrix $X \in \mathbb{R}^{n\times 2}$ is created by drawing $n$ samples from a multivariate normal distribution with two covariates. We build the matrix $Z$ with one covariate drawn from a uniform distribution. The matrix $W$ is created by assigning each sample to its row index and identifying its nearest neighbors based on these indices. Here, we considered 5 neighbors, but also tried with more (10 or 20) and the results were suggesting similar performances.

Then, the parameters $\mu$ and $\phi$ are computed from the matrices $X$ and $Z$ and the parameters $\beta$ and $\gamma$. The response matrix $Y$ is finally generated by drawing in the Dirichlet distribution of parameter $\alpha_i$, defined through $\mu$ and $\phi$, for each row $i$.

We perform 100 repetitions of the following: we create the data, and compute the bias of the estimated parameters. In regard to the non-spatial parameters, both models demonstrate similar behavior, with their bias, variance, and mean squared error being asymptotically unbiased. In the spatial models, we also observe the expected behavior, where the bias and mean squared error of the estimated $\hat{\rho}$ decrease as the number of samples increases. Interestingly, when $\hat{\rho}$ is biased (which occurs when the sample size $n$ is small), the bias is negative, suggesting that the model tends to underestimate the spatial correlation strength.

Then, the prediction accuracy of the models is assessed as follow. For each value of $\rho$, we create the test set by generating 1000 new data points using the true parameters $\beta^*$. On this test data, the true value $\mu^*$ is computed. Then, we compute $\hat{\mu}$ using the parameters $\hat{\beta}$ estimated from the $n = 1000$ simulation. To evaluate the difference between $\mu^*$ and each $\hat{\mu}$,

metrics such as $R^2$, RMSE, cross-entropy and cosine similarity are utilized.

We observe that for a low spatial correlation ($\rho = 0.1$), both models perform equally well. However, as the spatial correlation increases ($\rho = 0.5$ and $\rho = 0.9$), the performances of the non-spatial models decrease, and they are outperformed by the spatial model across all metrics. Actually, the best performances of the spatial model are observed under a moderately high spatial correlation ($\rho = 0.5$). This suggests that the performances of the spatial model are optimal around a moderate level of spatial dependence, but decline at low or extremely high spatial correlation levels.

## 3.2 Real datasets

### 3.2.1 Arctic lake

The Arctic Lake dataset (Coakley et Rust 1968) provides data for $n = 39$ sediment samples taken at various water depths in an Arctic lake. The label is compositional because it correspond to the percentage of sand, silt, and clay in the samples. Here, the goal is to analyze the influence of the water depth on the composition of the sediment samples. Here, we propose two regression models: one with a single predictor variable (the depth) with an intercept term, and another model with an intercept term, the depth variable, and its squared value.

The Leave One Out Cross Validation (LOOCV) strategy is particularly adapted to the small size of the dataset. We iteratively exclude one sample (the $k$-th sample) from the dataset, and use the remaining samples to estimate the parameters of the models. We then compute the predicted odd values $\hat{\mu}_k$ for the excluded sample, and evaluate its proximity to the true compositional label using metrics such as $R^2$, RMSE, cross-entropy and cosine similarity.

The results, not detailed here in the seek of simplicity, suggest that the utilization of spatial information leads to slight improvements in model performance. However, in terms of variability, the difference may not be statistically significant. This lack of significance could be attributed to the spatial information being derived solely from the depth variable, resulting in the absence of any new information being introduced. Instead, the data is essentially replicated in a different manner. What would be interesting, for instance, would be to have the exact spatial location of the different samples to use them to build the spatial weight matrix $W$.

### 3.2.2 Elections

The elections dataset present the votes at the French departmental election of 2015 in the Occitanie region (Goulard et al. 2017), for $n = 207$ cantons. For each canton, the voting distribution (initially between 15 political parties) is categorized into three major political movements: left, right, and extreme right. In our study, we utilized 25 distinct social indicators as features, including age categories, employment fields, and education level, among

others. Initially, the dataset consisted of 283 cantons, but any cantons where one of the classes was not present were removed. This resulted in the exclusion of 76 points, which represents 27% of the data.

The spatial weights matrix $W$ is computed based on the geographic proximity of each canton's center. Two cases are considered: in the first case, the contiguity-based, we consider the 5 nearest neighboring cantons, determined by their center-to-center distances. In the second case, the distance-based, the inverse of the distance between each canton and the others is considered, with a cut-off at a certain value that minimizes the average number of neighbors and to ensure that each canton has at least one neighbor. This cut-off gives 12 neighbors on average.

The matrix $Z$ is defined as a sole intercept, as this choice yielded the best results compared to the case where $Z$ was a copy of the features matrix $X$.

For the three models (non-spatial and the two spatials), we use the maximum likelihood estimator to retrieve the parameters and compute the performance with our usual metrics. Results are reported in Table 1. The estimated spatial correlation coefficient $\hat{\rho}$ is 0.97 (resp. 0.91) with the distance-based (resp. contiguity-based) matrix.

Table 1: Scores for the Dirichlet models on Elections dataset.

| **Model** | $\boldsymbol{R^2}$ | **RMSE** | **Cross-entropy** | **AIC** | **Cos similarity** |
|---|---|---|---|---|---|
| No spatial | 0.487 | 0.080 | 1.048 | -862.1 | 0.975 |
| Spatial (contiguity) | 0.582 | 0.072 | 1.042 | -947.4 | 0.979 |
| Spatial (distance) | **0.602** | **0.070** | **1.041** | **-965.1** | **0.980** |

The spatial models perform better than the non-spatial model across all evaluation metrics, excepted for the AIC which is slightly better for the non-spatial model. Besides, the distance-based spatial model performs slightly better than the contiguity-based. Additionally, we attempted to make predictions using our model through a 10-fold cross-validation technique, where 90% of the data were used to estimate the model parameters, while the remaining 10% (corresponding to 21 values) were reserved for testing the model's performance. However, we observed extremely poor performance on the test set, indicating that the spatial model is highly sensitive to missing values. This could be explained by the fact that 27% of the initial data was already missing, and removing more data might have rendered the spatial information irrelevant.

# 4 Conclusion

Our study demonstrates that incorporating spatial dependencies in a Dirichlet model increases the performances of the model. Our results from the real-life datasets reveal that a distance-based spatial weight matrix tends to yield better results compared to a contiguity-based matrix. These results underscore the potential advantages of spatial modeling, especially in scenarios where the Dirichlet distribution is well-suited to the data.

The results obtained from the synthetic dataset provide some insights into the behavior of the SAR Dirichlet model. While the spatial model outperforms the non-spatial model in spatially correlated data, the spatial model does not perform optimally under extremely high spatial correlation and provides better results when the spatial correlation is moderate ($\rho = 0.5$).

Overall, our study highlights the importance of considering spatial information when it provides meaningful additional context, as it can significantly enhance the model's effectiveness. It also emphasizes the potential impact of missing data, which should be carefully addressed to avoid adverse effects on model performance.

# Bibliographie


Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139-160.

Anselin, L. (2013). *Spatial econometrics: methods and models* (Vol. 4). Springer Science & Business Media.

Cliff, A. D., et Ord, K. (1970). Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography*, 46(sup1), 269-292.

Coakley, J. P., et Rust, B. R. (1968). Sedimentation in an Arctic lake. *Journal of Sedimentary Research*, 38(4), 1290-1300.

Goulard, M., Laurent, T., et Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, 12(2-3), 304-325.

Krisztin, T., Piribauer, P., et Wögerer, M. (2022). A spatial multinomial logit model for analysing urban expansion. *Spatial Economic Analysis*, 17(2), 223-244.

Leininger, T. J., Gelfand, A. E., Allen, J. M., et Silander, J. A. (2013). Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 314-334.

Maier, M. (2014). DirichletReg: Dirichlet regression for compositional data in R.

Nguyen, T. H. A., Thomas-Agnan, C., Laurent, T., et Ruiz-Gazen, A. (2021). A simultaneous spatial autoregressive model for compositional data.*Spatial Economic Analysis*, 16(2), 161-175.

Pirzamanbein, B., Lindström, J., Poska, A., et Gaillard, M. J. (2018). Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *Spatial statistics*, 24, 14-31.

# Réduction de la dimension sur données de distribution

Camille Mondon [1,*] & Anne Ruiz-Gazen [2,*] & Christine Thomas-Agnan [3,*]

[1] *Toulouse School of Economics, France,* camille.mondon@tse-fr.eu
[2] *Toulouse School of Economics, France,* anne.ruiz-gazen@tse-fr.eu
[3] *Toulouse School of Economics, France,* christine.thomas@tse-fr.eu

**Résumé.** Les données de distribution sont une généralisation continue en dimension infinie des données de composition. Elles sont souvent observées sous l'une des deux formes suivantes : valeurs non agrégées échantillonnées à partir de chaque loi ou valeurs agrégées sous forme d'histogrammes. Ces données sont généralement lissées afin de pouvoir appliquer des méthodes de réduction de la dimension. Pour les données non agrégées, nous proposons d'utiliser une méthode de lissage des échantillons maximisant une log-vraisemblance pénalisée. Nous obtenons des coefficients dans une base de splines sur lesquels nous appliquons la méthode *Invariant Coordinate Selection* multivariée pour réduire la dimension.

**Mots-clés.** Densités, données de composition, données fonctionnelles, invariant coordinate selection, lois elliptiques, splines de lissage.

**Abstract.** Distributional data are a continuous infinite-dimensional generalisation of compositional data. They are often observed in one of two forms : unaggregated values sampled from each distribution or data aggregated into histograms. A pre-processing smoothing step is usually applied before proceeding to dimension reduction techniques. For unaggregated data, we propose to use a sample smoothing method that maximises a penalized log-likelihood, in order to obtain spline coefficients on which we apply the multivariate Invariant Coordinate Selection method.

**Keywords.** Compositional data, densities, elliptical distributions, functional data, invariant coordinate selection, smoothing splines.

# Introduction

Le travail proposé porte sur l'élaboration et l'étude de méthodes statistiques pour des échantillons de variables aléatoires à valeurs dans un espace de densités de probabilités. L'analyse statistique de ce type de données, appelées données de distribution, est une démarche relativement récente qui connaît un essor notable en raison du volume et de la complexité croissante des données (Brito et Dias, 2022). De telles données sont présentes en sciences sociales avec les distributions d'âge ou de revenus, en climatologie avec les distributions de température, en géologie avec la composition géochimique d'échantillons de sols.

Il s'agit donc de variables fonctionnelles satisfaisant des contraintes non linéaires (positivité et intégrale égale à 1). Tout comme Petersen et al. (2022), on s'intéressera principalement au cas de densités de variables aléatoires absolument continues par rapport à la

mesure de Lebesgue. Une première idée pour développer de telles méthodes est d'utiliser une méthode existante pour données fonctionnelles et de la forcer à respecter les contraintes. Nous nous intéressons plutôt à des méthodes qui incorporent les contraintes dès le départ.

Petersen et al. (2022) présentent plusieurs approches, notamment des méthodes fondées sur une structure de variété pour l'espace de mesures de probabilités (Bigot et al., 2017), ou encore des méthodes reposant sur la notion d'espace de Bayes défini par Egozcue et al. (2006).

Nous nous concentrerons principalement sur la structure des espaces de Bayes en s'inspirant de ce qui existe déjà pour des variables discrètes avec un nombre fini de modalités. Dans ce cas, on peut considérer que les vecteurs de fréquences de ces densités sont des éléments d'un simplexe. On peut alors utiliser l'analyse des données de composition par la méthode des log-ratios introduite par Aitchison (1982).

Pour des variables continues, Van Den Boogaart et al. (2014) développent l'approche infini-dimensionnelle en construisant l'espace de Bayes $\mathcal{B}^2([a,b])$ des fonctions de densité (par rapport à la mesure de Lebesgue) et en le munissant d'une structure d'espace de Hilbert inspirée de la géométrie de Aitchison.

Hron et al. (2016) adaptent l'analyse en composantes principales (ACP) aux données de distribution agrégées sous forme d'histogrammes. Au cours de l'étape de pré-traitement, les histogrammes sont transformés en densités par un lissage de type moindre carrés utilisant des bases de fonctions splines qui appartiennent à un espace fonctionnel de dimension finie, d'après Machalová et al. (2016). Ensuite, ils construisent l'opérateur de covariance des transformées de rapport logarithmique d'après Van Den Boogaart et al. (2014), et résolvent le problème des valeurs propres et des vecteurs propres de l'ACP à l'aide des coefficients de la base de splines.

Tyler et al. (2009) présentent la méthode *Invariant Coordinate Selection* (ICS) comme une généralisation de l'ACP basée sur la diagonalisation jointe de deux matrices de dispersion. La méthode ICS transforme les données pour révéler le défaut d'ellipticité d'une distribution multivariée et peut être utilisée pour la détection des valeurs atypiques, comme le font Archimbaud et al. (2022) pour les données fonctionnelles, ou Ruiz-Gazen et al. (2023) pour les données de composition.

Nous proposons d'adapter la méthode ICS aux données de distribution de type échantillons agrégés et non agrégés. Pour des données agrégées en histogrammes, nous pouvons appliquer la méthode ICS sur données de composition définie par Ruiz-Gazen et al. (2023). Pour des échantillons non agrégés, nous proposons une approche différente présentée dans la section suivante.

# 1 Estimation des densités à partir d'échantillons non agrégés

Considérons $n$ échantillons $(x_k^i)_{1\leq k\leq N_i}$ d'observations dans un intervalle $[a,b]$, générés par des densités $f_i, 1\leq i\leq n$. Nous appliquons d'abord une étape de pré-traitement pour estimer directement chaque densité à l'aide de la méthode de maximisation de la log-vraisemblance pénalisée décrite par Silverman (1982) :

$$\hat{f}_i \in \underset{f\in\mathcal{B}^2([a,b])}{\operatorname{argmax}} \sum_{k=1}^{N_i} \log f(x_k^i) - \lambda \int_a^b (L(\log f)(x))^2 \mathrm{d}x$$

où $L$ est un opérateur différentiel, typiquement la dérivée troisième.

Si on restreint ce problème d'optimisation aux densités dont le logarithme appartient à un espace de splines de dimension finie, on peut le résoudre pour obtenir les coefficients des $\log \hat{f}_i$ sur une base de splines.

# 2 ICS pour données de distribution

Pour une variable aléatoire $X$ à valeurs dans $\mathbb{R}^p$, la méthode ICS revient à résoudre le problème de diagonalisation jointe suivant : trouver $H(X)\in\mathbb{R}^{p\times p}$ inversible telle que

$$H(X)^\top S_1(X) H(X) = I_p \text{ et } H(X)^\top S_2(X) H(X) = \Lambda(X)$$

ou de manière équivalente, diagonaliser

$$S_1(X)^{-1} S_2(X) = H(X)\Lambda(X)H(X)^{-1}$$

où $S_1(X)^{1/2}H(X)$ est une matrice orthogonale.

Suite au pré-traitement par maximisation de la log-vraisemblance pénalisée, on se restreint à un espace de dimension finie dans lequel on peut calculer des matrices de dispersion empiriques $\hat{S}_1$ et $\hat{S}_2$ des coefficients des log-densités, et leur appliquer la méthode ICS multivariée.

Dans la présentation, nous comparerons la méthode ICS distributionnelle à la méthode ICS compositionnelle appliquée aux histogrammes. La méthode ICS sur les histogrammes repose fortement sur le choix des classes, nécessite un traitement spécifique pour les classes vides et néglige l'ordre des classes, tandis que la méthode ICS sur données de distribution repose uniquement sur les nœuds et l'ordre des splines. Nous illustrerons les deux méthodes sur des données climatiques.

# Bibliographie


Aitchison, J. (1982). « The Statistical Analysis of Compositional Data ». *Journal of the Royal Statistical Society : Series B (Methodological)*, *44*(2), 139-160.

Archimbaud, A., Boulfani, F., Gendre, X., Nordhausen, K., Ruiz-Gazen, A., & Virta, J. (2022). « ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control ». *Econometrics and Statistics*.

Bigot, J., Gouet, R., Klein, T., & López, A. (2017). « Geodesic PCA in the Wasserstein space by convex PCA ». *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *53*(1), 1-26.

Brito, P., & Dias, S. (2022, avril). *Analysis of Distributional Data*. Chapman ; Hall/CRC.

Egozcue, J. J., Díaz–Barrero, J. L., & Pawlowsky–Glahn, V. (2006). « Hilbert Space of Probability Density Functions Based on Aitchison Geometry ». *Acta Mathematica Sinica*, *22*(4), 1175-1182.

Hron, K., Menafoglio, A., Templ, M., Hruzová, K., & Filzmoser, P. (2016). « Simplicial principal component analysis for density functions in Bayes spaces ». *Computational Statistics & Data Analysis*, *94*, 330-350.

Machalová, J., Hron, K., & Monti, G. (2016). « Preprocessing of centred logratio transformed density functions using smoothing splines ». *Journal of Applied Statistics*, *43*(8), 1419-1435.

Petersen, A., Zhang, C., & Kokoszka, P. (2022). « Modeling Probability Density Functions as Data Objects ». *Econometrics and Statistics*, *21*, 159-178.

Ruiz-Gazen, A., Thomas-Agnan, C., Laurent, T., & Mondon, C. (2023). « Detecting Outliers in Compositional Data Using Invariant Coordinate Selection ». In M. Yi & K. Nordhausen (Éd.), *Robust and Multivariate Statistical Methods : Festschrift in Honor of David E. Tyler* (p. 197-224). Springer International Publishing.

Silverman, B. W. (1982). « On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method ». *The Annals of Statistics*, *10*(3), 795-810.

Tyler, D. E., Critchley, F., Dümbgen, L., & Oja, H. (2009). « Invariant co-ordinate selection ». *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *71*(3), 549-592.

Van Den Boogaart, K. G., Egozcue, J. J., & Pawlowsky-Glahn, V. (2014). « Bayes Hilbert Spaces ». *Australian & New Zealand Journal of Statistics*, *56*(2), 171-194.

# Réseaux de neurones 2

# Domain Adaptation of Time Series through Optimal Transport and Temporal Alignment

François Painblanc[1], Laetitia Chapel[3], Nicolas Courty[2], Chloé Friguet[2], Charlotte Pelletier[2], Romain Tavenard[1]

[1] *Université Rennes 2, LETG, IRISA, Rennes, France*
[2] *Université Bretagne Sud, IRISA, UMR CNRS 6074, Vannes, France*
[3] *Institut Agro Rennes-Angers, IRISA, Rennes, France*

**Résumé.** De grandes quantités de données non étiquetées sont souvent disponibles, mais l'étape d'annotation est généralement une tâche fastidieuse et/ou coûteuse. En apprentissage non supervisé, l'adaptation de domaine peut résoudre ce problème en exploitant les étiquettes d'un domaine source pour classifier des données d'un domaine cible, similaire mais différent. Dans le cas de séries temporelles, des défis supplémentaires surviennent, notamment en raison des décalages temporels potentiels qui s'ajoutent aux décalages de distribution entre les domaines.

Nous présentons une méthode dénommée Match-And-Deform (MAD) qui vise à relever ces défis en identifiant les correspondances entre les séries temporelles des domaines source et cible tout en tenant compte des distorsions temporelles. Le problème d'optimisation associé aligne simultanément (1) les séries en optimisant un coût de transport optimal et (2) les temps à l'aide de *dynamic time warping*. Intégré dans un réseau de neurones profond, MAD permet l'apprentissage de nouvelles représentations des séries temporelles, alignant les domaines et améliorant le pouvoir discriminant du réseau.

La méthode est évaluée empiriquement sur des données de référence et des données réelles de télédétection, démontrant l'efficacité de MAD: les séries sont appariées de façon pertinentes et les décalages temporels sont estimés avec précision. Des performances de classification comparables ou supérieures aux stratégies d'adaptation de domaine de séries temporelles profondes l'état de l'art sont également obtenues.

**Cet article est issu de Painblanc et al. [2023]. Le code et le jeu de données sont disponibles publiquement:** `https://github.com/rtavenar/MatchAndDeform`

**Mots-clés.** adaptation de domaines, séries temporelles, transport optimal, dynamic time warping

**Abstract.** Large amounts of unlabeled data are often available, and the annotation step is usually a tedious and/or costly task. Unsupervised domain adaptation can address this issue by leveraging labels from a source domain to classify data from a related, yet different, target domain. When dealing with time series data, additional challenges arise due to potential temporal shifts alongside the feature distribution shift.

We introduce the Match-And-Deform (MAD) approach to address these challenges. MAD aims at identifying matching between source and target time series while taking into account temporal distortions. To achieve this, the associated optimization problem simultaneously (1)

aligns the series using an optimal transport loss and (2) adjusts the timestamps using dynamic time warping. When integrated into a deep neural network, MAD facilitates the learning of new representations of time series that align the domains and enhance the network's discriminative power.

Empirical evaluations conducted on benchmark datasets and real remote sensing data demonstrate MAD's effectiveness. These numerical experiments show meaningful sample-to-sample matching and accurately estimates time shifts. Comparable or superior classification performance compared to state-of-the-art deep time series domain adaptation strategies are also achieved.

**This paper is adapted from Painblanc et al. [2023]. Code and data are publicly available:** `https://github.com/rtavenar/MatchAndDeform`

**Keywords.** domain adaptation, time series, optimal transport, dynamic time warping

# 1 Introduction

A standard assumption in machine learning is that the training and the test data are drawn from the same distribution. When this assumption is not met, trained models often have degraded performances because of their poor generalization ability. Domain Adaptation (DA) addresses this challenge by considering the generalization problem when there exists a distributional shift, allowing for improving task efficiency on the target domain through the use of comprehensive information from a source domain.

In this work, we consider DA issue for time series data. Our objective is to classify time series from an unlabelled target dataset $(\mathbf{X}') \in \mathbb{R}^{n' \times T' \times q}$ using a labelled source dataset $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{n \times T \times q} \times \mathcal{C}$. This setting can be yield for example by crop-type mapping from remote sensing data (miniTimeMatch dataset), where domains correspond to different geographical areas and class-level temporal shifts are observed, as illustrated on Fig.1.

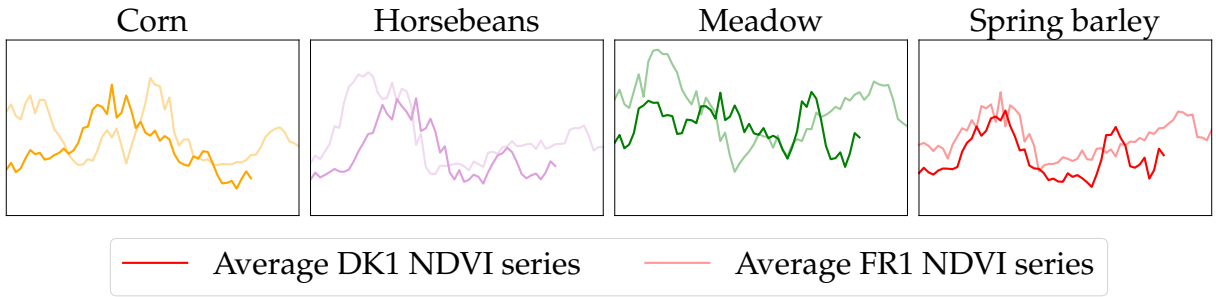



Figure 1: Illustration of temporal shift between averaged NDVI time series for 2 domains, for different types of crops (miniTimeMatch data)

The DA literature primarily focuses on addressing distribution shifts through alignment or common representation space approaches. In unsupervised DA frameworks, training on

source domain data leverages this shared representation for improved performance on the target domain. The standard adversarial training aims to induce domain-invariant representations in deep neural networks. Extention to time series can be done with specialized architectures such as convolutional layers (CoDATS, Wilson et al. [2020]). However, the time dimension vanishes with the use of pooled features. Optimal Transport (OT, Peyré and Cuturi [2019]) emerges as a powerful tool in both unsupervised and semi-supervised DA, deriving efficient solutions for assessing distribution shifts and deep neural network losses that capture domain dissimilarities. However, current OT-based DA methods do not encode any temporal coherence for time series data analysis.

In the following, we first introduce basics from OT and time series alignment, and define our optimisation problem to deal with DA for time series. Empirical evaluations are then conducted on benchmark datasets and real remote sensing data to demonstrate the effectiveness of the proposed approach.

# 2 Domain Adaptation for time series with Optimal Transport and Temporal Alignment

## 2.1 Background

The optimisation problem for comparing two objects (either time series or distributions) $\mathbf{x}$ and $\mathbf{x}'$ can be stated as:

$$J\Big(\mathbf{C}(\mathbf{x},\mathbf{x}'),\Pi\Big) = \underset{\boldsymbol{\pi}\in\Pi}{\arg\min} \Big\langle \mathbf{C}(\mathbf{x},\mathbf{x}'),\boldsymbol{\pi}\Big\rangle \ , \tag{1}$$

in which $\Pi$ is a set of admissible couplings. A coupling will either be a temporal alignment if $\mathbf{x}$ and $\mathbf{x}'$ are time series or a matching between samples if they are distributions, with appropriate constraint sets (see Fig.2.1). The solution $J(\cdot,\cdot)$ of the optimization problem is called the optimal coupling matrix. The cost matrix $\mathbf{C}(\mathbf{x},\mathbf{x}') = \{d(x^i,x'^j)\}_{ij}$ stores distances $d(x^i,x'^j)$ between atomic elements $x^i$ and $x'^j$, respectively from $\mathbf{x}$ and $\mathbf{x}'$. Dynamic time warping (DTW, Sakoe and Chiba [1978]) and Optimal transport (OT) are both instances of the same general optimization problem (1) that (OT) defines a distance between two probability measures or (DTW) matches two (multivariate) time series, as illustrated on Fig.2.1.

**Optimal Transport**

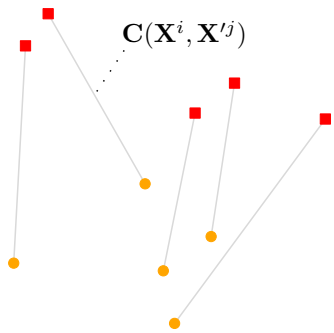
$$\underset{\boldsymbol{\gamma}\in\Gamma(\mathbf{w},\mathbf{w}')}{\arg\min} \left\langle \mathbf{C}(\mathbf{X},\mathbf{X}'), \boldsymbol{\gamma} \right\rangle$$

$\mathbf{X}, \mathbf{X}'$: Datasets with weights $\mathbf{w}, \mathbf{w}'$
$\Gamma(\mathbf{w}, \mathbf{w}')$: set of all couplings

**Dynamic Time Warping**

$$\underset{\boldsymbol{\pi}\in\mathcal{A}(T,T')}{\arg\min} \left\langle \mathbf{C}(\mathbf{x},\mathbf{x}'), \boldsymbol{\pi} \right\rangle$$

$\mathbf{x}, \mathbf{x}'$: Time series of lengths $T, T'$
$\mathcal{A}(T, T')$: set of all temporal alignments



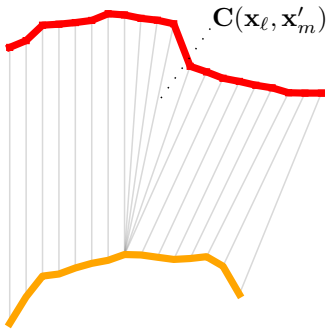




Figure 2: Illustration of the 2 optimisation problems: (left) OT defines a distance between two probability measures and (right) DTW matches two (multivariate) time series

## 2.2 Match-And-Deform

We introduce Match-And-Deform (MAD) that combines OT with DTW to achieve time series matching and timestamp alignment. In other words, MAD evaluates the feature distribution shift between domains up to a global temporal alignment. MAD jointly optimizes a global DTW alignment and an OT coupling to match two sets of time series. Let us therefore define MAD as:

$$\begin{aligned} \mathrm{MAD}(\mathbf{X},\mathbf{X}') &= \underset{\substack{\boldsymbol{\gamma}\in\Gamma(\mathbf{w},\mathbf{w}')\\ \boldsymbol{\pi}\in\mathcal{A}(T,T')}}{\arg\min} \langle \mathbf{L}(\mathbf{X},\mathbf{X}') \otimes \boldsymbol{\pi}, \boldsymbol{\gamma} \rangle \\ &= \underset{\substack{\boldsymbol{\gamma}\in\Gamma(\mathbf{w},\mathbf{w}')\\ \boldsymbol{\pi}\in\mathcal{A}(T,T')}}{\arg\min} \sum_{i,j} \sum_{\ell,m} d(x^i_\ell, x'^j_m) \pi_{\ell m} \gamma_{ij}. \end{aligned} \tag{2}$$

with, $\mathbf{L}(\mathbf{X}, \mathbf{X}')$ is a 4-dimensional tensor whose elements are $L^{i,j}_{\ell,m} = d(x^i_\ell, x'^j_m)$, with $d : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^+$ being a distance. $\otimes$ is the tensor-matrix multiplication. $\boldsymbol{\pi}$ is a global DTW alignment between timestamps and $\boldsymbol{\gamma}$ is a transport plan between samples from $\mathbf{X}$ and $\mathbf{X}'$.

This optimization problem can be further extended to the case of distinct DTW mappings for each class $c$ in the source data. This results in the following optimization problem, coined $|\mathcal{C}|$-MAD:

$$|\mathcal{C}|\text{-MAD}(\mathbf{X},\mathbf{X}',\mathbf{Y}) = \underset{\substack{\boldsymbol{\gamma}\in\Gamma(\mathbf{w},\mathbf{w}')\\ \forall c, \boldsymbol{\pi}^{(c)}\in\mathcal{A}(T,T')}}{\arg\min} \sum_{i,j} \sum_{\ell,m} L^{i,j}_{\ell,m} \pi^{(y^i)}_{\ell m} \gamma_{ij} \,. \tag{3}$$

In that case, $|\mathcal{C}|$ DTW alignments are involved, one for each class $c$. $\boldsymbol{\pi}^{(y^i)}$ denotes the DTW matrix associated to the class $y^i$ of $x^i$. This more flexible formulation allows adapting to different temporal distortions that might occur across classes.

The joint optimization problem introduced in Eq. (3) involves $|\mathcal{C}|$ finite sets of admissible DTW paths and a continuous space with linear constraints for the OT plan. We perform a Block Coordinate Descent (BCD) to optimize the corresponding loss.

Fig. 3 illustrates the general workflow of MAD.

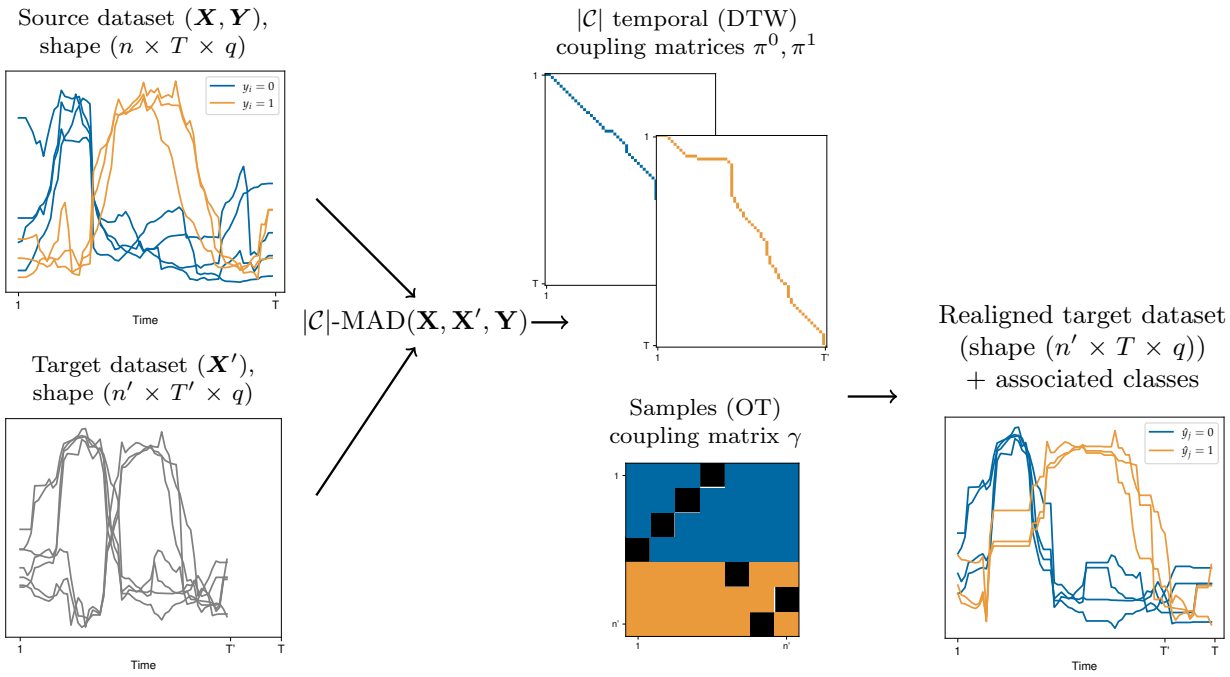



Figure 3: Match-And-Deform ($|\mathcal{C}|$-MAD) takes two time series datasets as inputs: a source (labelled) dataset and a target (unlabelled) dataset. It jointly computes an optimal transport (OT) coupling matrix $\boldsymbol{\gamma}$ and $|\mathcal{C}|$ class-wise dynamic time warping (DTW) paths $\{\boldsymbol{\pi}^{(c)}\}_{c \in \mathcal{C}}$. The OT cost is derived from the pairwise distances yielded by the DTW paths while the DTW cost is weighted by the OT plan. These outputs are then used to improve classification in the target dataset (figure from Painblanc et al. [2023])

## 2.3 Deep Domain Adaptation Loss

MAD can be used as a loss function in a neural network to learn a domain-invariant latent representation. Indeed, OT has been successfully used as a loss to measure the discrepancy between source and target domain samples embedded into a latent space. Similarly to DeepJDOT (Damodaran et al. [2018]), our proposal considers a deep unsupervised temporal DA model that relies on MAD or $|\mathcal{C}|$-MAD as a regularization loss function, as illustrated on Fig.4.

$$\mathcal{L}\Big(\mathbf{X}, \mathbf{Y}, \mathbf{X}', f_\theta, g_\Omega, \gamma, \{\pi^{(c)}\}_c\Big) = \frac{1}{n}\sum_i \overbrace{\mathcal{L}_s\Big(y^i, f_\theta(g_\Omega(\mathbf{x}^i))\Big)}^{\text{Cross-Entropy (CE) on source domain}} \tag{4}$$
$$+ \alpha \underbrace{\sum_{i,j}\sum_{\ell,m} d\left(g_\Omega(\mathbf{x}^i)_\ell, g_\Omega(\mathbf{x}'^j)_m\right) \pi^{(y^i)}_{\ell m} \gamma_{ij}}_{\substack{\text{Matching series under temporal alignment} \\ \text{(MAD cost on intermediate features)}}} + \beta \underbrace{\sum_{i,j} \mathcal{L}_t\Big(y^i, f_\theta(g_\Omega(\mathbf{x}'^j))\Big)\gamma_{ij}}_{\substack{\text{CE on target domain} \\ \text{transporting source labels}}}$$

Optimisation is computed over two groups of parameters: (i) the neural network parameters $\theta$ and $\Omega$ and (ii) MAD transport plan $\boldsymbol{\gamma}$ and DTW paths $\{\boldsymbol{\pi}^{(c)}\}_c$. Similar to what is done in Damodaran et al. [2018], we use an approximate optimization procedure that relies on stochastic gradients.

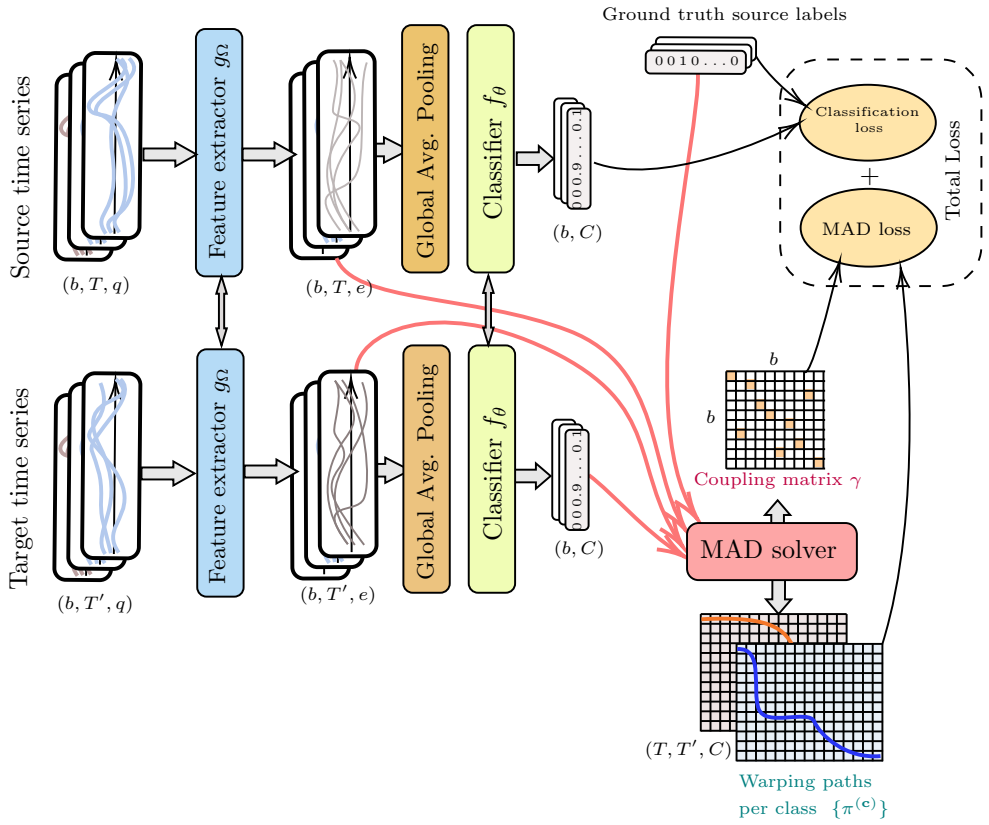



Figure 4: MAD backbone architecture and schematic view of the loss computation (figure from Painblanc et al. [2023])

# 3 Experimental Results on Remote Sensing Data

The use of MAD and $|\mathcal{C}|$-MAD as losses for neural domain adaptation is now assessed considering a real remote sensing dataset, for which there exists a known global temporal shift between the (classes of the) domains due to different weather conditions. The proposed approach has also been further studied on benchmarked datasets, results are not reported here.

miniTimeMatch dataset is a subsample of TimeMatch Nyborg et al. [2022], a crop-type mapping dataset of different geographical areas, with assumed temporal shifts. Pre-processing performed on the raw data is described in Painblanc et al. [2023]. It finally leads to a dataset of $28,858$ time series and 8 classes per domain, each being described by 10 features per timestamp. Both variant of MAD (considering a single DTW path *vs* one DTW path per class) are evaluated, in comparison with state-of-the-art method CoDATS-WS (Wilson et al. [2020]). Main results are reported in Tab.1: MAD and $|\mathcal{C}|$-MAD outperform CoDATS-WS in 4 out of the 5 DA problems, sometimes with an important improvement (see DK1 $\rightarrow$ FR1 for example). This illustrates the fact that MAD and $|\mathcal{C}|$-MAD are of prime interest when global or class-specific temporal deformations occur between domains (see Fig. 1).

| Problem | No adaptation | CoDATS-WS | MAD | $\|\mathcal{C}\|$-MAD | Target only |
|---|---|---|---|---|---|
| DK1 $\rightarrow$ FR1 | $69.2 \pm 1.3$ | $74.8 \pm 1.5$ | $\mathbf{88.4 \pm 0.4}$ | $88.3 \pm 0.9$ | $95.8 \pm 0.9$ |
| DK1 $\rightarrow$ FR2 | $62.2 \pm 3.5$ | $\mathbf{87.0 \pm 3.4}$ | $82.5 \pm 1.1$ | $81.0 \pm 1.1$ | $94.2 \pm 1.7$ |
| DK1 $\rightarrow$ AT1 | $73.9 \pm 0.2$ | $71.6 \pm 15.4$ | $\mathbf{93.1 \pm 1.2}$ | $92.3 \pm 2.2$ | $96.7 \pm 0.7$ |
| FR1 $\rightarrow$ DK1 | $61.9 \pm 5.2$ | $78.0 \pm 10.7$ | $\mathbf{88.2 \pm 0.3}$ | $\mathbf{88.2 \pm 0.5}$ | $96.2 \pm 0.3$ |
| FR1 $\rightarrow$ FR2 | $78.8 \pm 0.9$ | $82.1 \pm 8.2$ | $\mathbf{90.5 \pm 0.2}$ | $89.6 \pm 0.4$ | $94.2 \pm 1.7$ |
| Average | $69.2 \pm 2.2$ | $78.7 \pm 7.8$ | $\mathbf{88.5 \pm 0.6}$ | $87.9 \pm 1.0$ | $95.4 \pm 1.1$ |

Table 1: Mean and std classification accuracy over 3 repetitions (DK: Denmark, FR: France, AT: Austria)

# 4 Conclusion and perspectives

In this paper, we introduce Match-And-Deform (MAD) that combines optimal transport and dynamic time warping for time series domain adaptation in the presence of global time shifts. We furthermore embed MAD as a regularization loss in a neural domain adaptation setting and evaluate its performance in different settings: MAD reaches better performance than state-of-the-art strategies thanks to its ability to capture temporal shifts.

Nevertheless, inter-domain class balance is an implicit OT hypothesis. Extension of MAD could alleviate this OT assumption by using unbalanced optimal transport. An application of MAD could be to consider an estimate the quality of missing values imputation with MAD score.

# References


Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 447–463, 2018.

Joachim Nyborg, Charlotte Pelletier, Sébastien Lefèvre, and Ira Assent. Timematch: Unsupervised cross-region adaptation by temporal shift estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:301–313, 2022.

François Painblanc, Laetitia Chapel, Nicolas Courty, Chloé Friguet, Charlotte Pelletier, and Romain Tavenard. Match-And-Deform: Time Series Domain Adaptation through Optimal Transport and Temporal Alignment. In *ECML PKDD 2023*, Torino, Italy, September 2023. URL `https://hal.science/hal-04189149`.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1): 43–49, 1978.

Garrett Wilson, Janardhan Rao Doppa, and Diane J Cook. Multi-source deep domain adaptation with weak supervision for time-series sensor data. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1768–1778, 2020.

# Analyse non asymptotique des algorithmes stochastiques adaptatifs biaisés

Sobihan Surendran [1,2], Adeline Fermanian [2], Antoine Godichon-Baggioni [1]
Sylvain Le Corff [1]

[1] *LPSM, Sorbonne Université, France*
{*sobihan.surendran, antoine.godichon_baggioni, sylvain.le_corff*}*@sorbonne-universite.fr*
[2] *LOPF, Califrais, Machine Learning Lab, Paris, France*
{*sobihan.surendran, adeline.fermanian*}*@califrais.fr*

**Résumé.** La Descente de Gradient Stochastique (SGD) avec pas adaptatifs est désormais largement utilisée, en particulier pour l'apprentissage des réseaux neuronaux profonds. Cependant, la plupart des résultats théoriques supposent l'accès à des estimateurs du gradient non biaisés, ce qui n'est pas le cas dans de nombreuses applications récentes en apprentissage profond et en apprentissage par renforcement utilisant des méthodes de Monte Carlo. Nous proposons dans cette présentation une analyse non asymptotique de l'algorithme SGD utilisant des estimateurs biasés du gradient ainsi que des pas adaptatifs pour les fonctions convexes et non convexes. Notre étude intègre un biais dépendant du temps et met l'accent sur l'importance de contrôler le biais et l'erreur quadratique moyenne de l'estimateur du gradient. En particulier, nous établissons que les algorithmes Adagrad et RMSProp avec gradients biaisés convergent vers des points critiques à une vitesse de convergence similaire aux résultats existants dans la littérature pour le cadre non biaisé. Enfin, nous fournissons des résultats expérimentaux utilisant des Autoencodeurs Variationnels qui illustrent nos résultats de convergence et montrent comment l'effet du biais peut être réduit par un réglage approprié des hyperparamètres.

**Mots-clés.** Optimisation Stochastique, Approximation Stochastique Biaisée, Méthodes de Monte Carlo, Autoencodeurs Variationnels

**Abstract.** Stochastic Gradient Descent (SGD) with adaptive steps is now widely used for training deep neural networks. Most theoretical results assume access to unbiased gradient estimators, which is not the case in several recent deep learning and reinforcement learning applications that use Monte Carlo methods. We provide a comprehensive non-asymptotic analysis of SGD with biased gradients and adaptive steps for convex and non-convex smooth functions. Our study incorporates time-dependent bias and emphasizes the importance of controlling the bias and Mean Squared Error of the gradient estimator. In particular, we establish that Adagrad and RMSProp with biased gradients converge to critical points for smooth non-convex functions at a rate similar to existing results in the literature for the unbiased case. Finally, we provide experimental results using Variational Autoenconders (VAE) that illustrate our convergence results and show how the effect of bias can be reduced by appropriate hyperparameter tuning.

**Keywords.** Stochastic Optimization, Biased Stochastic Approximation, Monte Carlo Methods, Variational Autoenconders

# 1 Introduction

Les algorithmes de Descente de Gradient Stochastique (SGD) sont des méthodes classiques pour entraîner des modèles statistiques basés sur des architectures profondes. Considérons le problème d'optimisation :

$$\theta_* \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} V(\theta), \tag{1}$$

où $V$ est la fonction objectif, supposée différentiable. La descente de gradient stochastique est définie par $\theta_0 \in \mathbb{R}^d$ et pour tout $n \geq 1$,

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \widehat{\nabla V}(\theta_n),$$

où $\widehat{\nabla V}(\theta_n)$ est un estimateur du gradient de $V$. En apprentissage profond, la stochasticité émerge par exemple par l'utilisation de mini-batchs, étant donné qu'il n'est pas possible de calculer les gradients sur l'ensemble des données. Bien que ces algorithmes aient fait l'objet de nombreuses études, tant sur le plan théorique que pratique [Bottou et al., 2018], de nombreuses questions restent ouvertes. En particulier, la plupart des résultats théoriques reposent sur l'hypothèse que l'estimateur de gradient est non biaisé. Cependant, cette hypothèse n'est pas vérifiée pour de nombreuses applications courantes. Par exemple, dans les algorithmes d'apprentissage par renforcement, les estimateurs du gradient sont souvent obtenus à partir d'une chaîne de Markov, ce qui conduit à des estimateurs biaisés [Sun et al., 2018, Doan et al., 2020]. D'autres exemples de gradients biaisés peuvent être trouvés dans le domaine des modèles génératifs utilisant des méthodes de Monte Carlo par chaînes de Markov (MCMC) et des méthodes de Monte Carlo séquentielles (SMC) [Gloaguen et al., 2022, Cardoso et al., 2023]. En particulier, l'approche IWAE proposée par Burda et al. [2015], qui est une variante de l'Autoencodeur variationnel standard (VAE) [Kingma and Welling, 2013], produit des estimateurs de gradient biaisés.

Dans les applications pratiques, l'algorithme SGD standard rencontre des difficultés pour calibrer les suites de pas. Par conséquent, les variantes modernes utilisent des pas adaptatifs basés sur les gradients ou la matrice Hessienne pour éviter les points selles et traiter les problèmes mal conditionnés. L'idée des pas adaptatifs a d'abord été proposée dans la littérature portant sur l'apprentissage en ligne par Auer et al. [2002] et ensuite adoptée pour l'optimisation stochastique, avec l'algorithme Adagrad de Duchi et al. [2011].

À notre connaissance, aucune analyse non asymptotique tenant compte à la fois de l'estimateur biaisé du gradient et des pas adaptatifs n'a été menée à ce jour. Nous présentons des garanties de convergence pour l'algorithme SGD avec des gradients biaisés et des pas adaptatifs, sous des hypothèses faibles portant sur le biais et l'erreur quadratique moyenne de l'estimateur. Dans de nombreux scénarios, il est en effet possible de contrôler ces quantités, comme cela a été récemment montré, par exemple, pour l'Échantillonnage d'Importance et les méthodes de Monte Carlo Séquentielles. En particulier, nous établissons qu'Adagrad et RMSProp avec un gradient biaisé convergent vers un point critique pour les fonctions non convexes avec une vitesse de convergence en $\mathcal{O}(\log n/\sqrt{n} + b_n)$, où $b_n$ est lié au biais à l'itération $n$. Pour les fonctions convexes, nous obtenons une vitesse de convergence

amélioréoe en $\mathcal{O}(1/\sqrt{n} + b_n)$. Nos résultats théoriques nous fournissent des procédures de réglage des hyperparamètres pour éliminer efficacement le terme de biais, ce qui se traduit par de meilleures vitesses de convergence de l'ordre de $\mathcal{O}(\log n/\sqrt{n})$ et $\mathcal{O}(1/\sqrt{n})$ respectivement.

# 2 Algorithmes Stochastiques Adaptatifs Biasés

## 2.1 Cadre

Considérons le problème d'optimisation (1) et l'algorithme de gradient stochastique à pas adaptatifs biaisés suivant : $\theta_0 \in \mathbb{R}^d$ et pour tout $n \geq 0$,

$$\theta_{n+1} = \theta_n - \gamma_{n+1} A_n H_{\theta_n}\left(X_{n+1}\right) \ , \tag{2}$$

où $\gamma_{n+1} > 0$ et $A_n$ est une suite de matrices symétriques et définies positives. Soit $(\mathcal{F}_n)_{n\geq 0}$ la filtration générée par les variables aléatoires $(\theta_0, \{X_k\}_{k\leq n})$ et supposons que pour tout $n \geq 0, A_n$ est $\mathcal{F}_n$-mesurable. Dans un contexte d'estimations de gradient biaisées, le choix de

$$A_n = \left[\delta I_d + \Big(\frac{1}{n}\sum_{k=0}^{n-1} H_{\theta_k}\left(X_{k+1}\right) H_{\theta_k}\left(X_{k+1}\right)^\top\Big)\right]^{-1/2}$$

peut être assimilé à l'algorithme full Adagrad [Duchi et al., 2011]. Cependant, calculer la racine carrée de l'inverse devient coûteux en grande dimension, donc en pratique, Adagrad est souvent utilisé avec des matrices diagonales. En notant $\mathrm{Diag}(A)$ la matrice formée avec les termes diagonaux de $A$ et en annulant tous les autres termes, Adagrad est défini dans notre contexte par :

$$A_n = \left[\delta I_d + \mathrm{Diag}(\bar{H}_n(X_{0:n}, \theta_{0:n-1}))\right]^{-1/2}, \tag{3}$$

où

$$\bar{H}_n(X_{0:n}, \theta_{0:n-1}) = \frac{1}{n}\sum_{k=0}^{n-1} H_{\theta_k}(X_{k+1}) H_{\theta_k}(X_{k+1})^\top.$$

Dans RMSProp [Tieleman et al., 2012], $\bar{H}_n(X_{0:n}, \theta_{0:n-1})$ dans (3) est une moyenne mobile exponentielle des carrés des gradients, définie par :

$$(1-\rho)\sum_{k=0}^{n-1} \rho^{n-k} H_{\theta_k}(X_{k+1}) H_{\theta_k}(X_{k+1})^\top,$$

où $\rho$ est le paramètre de la moyenne mobile. De plus, lorsque $A_n$ est une estimation récursive de l'inverse de la Hessienne, cela correspond à l'algorithme Newton Stochastique [Boyer and Godichon-Baggioni, 2023].

## 2.2 Hypothèses

Nous énonçons ci-dessous les principales hypothèses nécessaires à nos résultats théoriques.

**Hypothèse 1.** *La fonction objectif $V$ est convexe et il existe une constante $\mu > 0$ telle que :*

$$2\mu \left(V(\theta) - V\left(\theta^*\right)\right) \leq \|\nabla V(\theta)\|^2, \quad \forall \theta \in \mathbb{R}^d.$$

La deuxième condition de l'Hypothèse 1 correspond à la condition de Polyak-Łojasiewicz. De plus, l'Hypothèse 1 est une hypothèse plus faible par rapport à la forte convexité de la fonction.

**Hypothèse 2.** *Le gradient de la fonction objectif $V$ est Lipschitz. Pour tout $(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d$,*

$$\|\nabla V(\theta) - \nabla V\left(\theta'\right)\| \leq L \left\|\theta - \theta'\right\|.$$

Cette hypothèse est cruciale pour obtenir notre vitesse de convergence et est très courante [Moulines and Bach, 2011, Bottou et al., 2018].

**Hypothèse 3.** *Pour tout $n \in \mathbb{N}$, il existe $r_n \geq 0$ et $\sigma_n^2 \geq 0$ tels que :*

*(i) Biais :* $\left\|\mathbb{E}[H_{\theta_n}(X_{n+1}) \mid \mathcal{F}_n] - \nabla V(\theta_n)\right\| \leq r_n$.

*(ii) Erreur quadratique moyenne :*

$$\mathbb{E}\left[\left\|H_{\theta_n}\left(X_{n+1}\right) - \nabla V\left(\theta_n\right)\right\|^2 \mid \mathcal{F}_n\right] \leq \sigma_n^2.$$

Dans cette hypothèse, les suites $r_n$ et $\sigma_n^2$ contrôlent le biais et l'erreur quadratique moyenne sans aucune hypothèse spécifique sur leur dépendance par rapport à $n$, ce qui rend notre cadre très général. Cette hypothèse peut être vérifiée dans diverses applications utilisant des méthodes de Monte Carlo. Si $H_{\theta_n}\left(X_{n+1}\right)$ est un estimateur non biaisé du gradient, le deuxième point est équivalent à garantir que la variance du terme de bruit est bornée. Dans ce contexte, cette hypothèse est standard, voir par exemple Moulines and Bach [2011], Ghadimi and Lan [2013].

Nous considérons également une hypothèse supplémentaire sur $A_n$. Soit $\|A\|$ la norme spectrale d'une matrice $A$.

**Hypothèse 4.** *Pour tout $n \in \mathbb{N}$, il existe $\beta_n, \lambda_n > 0$ tels que :*

$$\|A_n\| := \lambda_{\max}(A_n) \leq \beta_{n+1} \quad \textit{et} \quad \lambda_{\min}\left(A_n\right) \geq \lambda_{n+1}.$$

Dans notre cadre, puisque $A_n$ est supposée être une matrice symétrique, la norme spectrale est égale à la plus grande valeur propre.

**Hypothèse 5.** *Le gradient est borné, c'est-à-dire qu'il existe $M \geq 0$ tel que pour tout $n \in \mathbb{N}$,*

$$\|H_{\theta_n}\left(X_{n+1}\right)\| \leq M.$$

Il est important de noter que sous l'Hypothèse 3, ceci équivaut à borner le gradient stochastique de la fonction objectif.

## 2.3 Résultats de Convergence

### 2.3.1 Cas Convexe

Dans cette section, nous étudions les résultats de convergence pour l'algorithme SGD avec des gradients biaisés et des pas adaptatifs dans le cas convexe. Nous donnons ci-dessous une version simplifiée de la borne que nous avons obtenue sur le risque sans aucune constante explicite.

**Théorème 2.1.** *Soit $\theta_n \in \mathbb{R}^d$ la n-ème itération de la récursion (2) et $\gamma_n = C_\gamma n^{-\gamma}$, $\beta_n = C_\beta n^{\beta}$, $\lambda_n = C_\lambda n^{-\lambda}$ avec $C_\gamma > 0$, $C_\beta > 0$ et $C_\lambda > 0$. Supposons que $\gamma, \beta, \lambda \geq 0$ et $\gamma + \lambda < 1$. Alors, sous les hypothèses $1-4$, nous avons :*

$$\mathbb{E}\left[V\left(\theta_n\right) - V(\theta^*)\right] = \mathcal{O}\left(n^{-\gamma+2\beta+\lambda} + b_n\right), \tag{4}$$

*où le terme de biais $b_n$ peut être constant ou décroissant. Dans le dernier cas, en écrivant $r_n = n^{-\alpha}$, nous avons :*

$$b_n = \mathcal{O}\left(n^{-2\alpha+2\beta+2\lambda}\right).$$

Le résultat obtenu est classique et montre le compromis entre un terme provenant des pas adaptatifs (avec une dépendance en $\gamma$, $\beta$, $\lambda$) et un terme $b_n$ qui dépend du contrôle du biais $r_n$. Pour minimiser le membre de droite de (4), nous aimerions avoir $\beta = \lambda = 0$. Cependant, cela nécessiterait des hypothèses beaucoup plus fortes. Par exemple, dans le cas d'Adagrad et de RMSProp, les gradients devraient être bornés. Dans le cas de SGD avec échantillonnage de coordonnées mais sans pas adaptatifs, le résultat analogue peut être trouvé dans Leluc and Portier [2022].

### 2.3.2 Cas non convexe à gradient Lipschitz

Dans le cas non convexe à gradient Lipschitz, nous avons obtenu la vitesse de convergence sans faire l'hypothèse d'un gradient borné, cependant, nous présentons uniquement les résultats pour le cas d'un gradient borné dans le cadre de l'Approximation Stochastique Adaptative Randomisée. Le théorème suivant fournit une borne en espérance sur le gradient de la fonction objectif $V$, ce qui est le meilleur résultat que nous puissions obtenir étant donné qu'aucune hypothèse n'est faite sur l'existence d'un minimum global de $V$.

**Théorème 2.2.** *Soient $\gamma_n = C_\gamma n^{-\gamma}, \beta_n = C_\beta n^{\beta}, \lambda_n = C_\lambda n^{-\lambda}$ avec $C_\gamma > 0, C_\beta > 0$, et $C_\lambda > 0$. Supposons que $\gamma, \beta, \lambda \geq 0$ et $\gamma + \lambda < 1$. Pour tout $n \geq 1$, soit $R \in \{0, \ldots, n\}$ une variable aléatoire uniformément distribuée. Alors, sous les hypothèses $2-4, 5$, nous avons :*

$$\mathbb{E}\left[\left\|\nabla V\left(\theta_R\right)\right\|^2\right] \leq \frac{2V^* + \alpha_{1,n} + LM^2\alpha_{2,n}}{\sqrt{n}},$$

*où $V^* = \mathbb{E}[V(\theta_0) - V(\theta^*)]$, $\alpha_{1,n} = \sum_{k=0}^{n} \gamma_{k+1}\beta_{k+1}^2 r_k^2/\lambda_{k+1}$ et $\alpha_{2,n} = \sum_{k=0}^{n} \gamma_{k+1}^2\beta_{k+1}^2$.*

### 2.3.3 Application à Adagrad et RMSProp

Dans cette section, nous fournissons l'analyse de convergence d'Adagrad et de RMSProp avec un estimateur de gradient biaisé.

**Corollaire 2.1.** *Soient $\gamma_n = c_\gamma n^{-1/2}$ et $A_n$ la matrice adaptative dans Adagrad ou RMSProp. Pour tout $n \geq 1$, soit $R \in \{0, \ldots, n\}$ une variable aléatoire uniformément distribuée. Alors, sous les Hypothèses $2, 3, 5$, nous avons :*

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \begin{cases} \mathcal{O}\left(n^{-2\alpha}\right) & \text{if } \alpha < 1/4\,, \\ \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right) & \text{if } \alpha \geq 1/4\,. \end{cases}$$

Le Corollaire 2.1 établit la vitesse de convergence d'Adagrad et de RMSProp avec un gradient biaisé vers un point critique pour les fonctions non convexes à gradient Lipschitz. Dans le cas d'un gradient non biaisé, nous obtenons la même borne que dans Zou et al. [2018], sous les mêmes hypothèses :

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right).$$

Si le biais est de l'ordre $\mathcal{O}(n^{-1/4})$, l'algorithme atteint le même vitesse de convergence que dans le cas d'un gradient non biaisé.

# 3 Illustrations numériques

Dans cette section, nous illustrons nos résultats théoriques dans le contexte des Variational Autoencoders (VAE) profonds. Dans les modèles génératifs, l'objectif est de maximiser la vraisemblance marginale définie comme suit :

$$\log p_\theta(x) = \log \mathbb{E}_{p_\theta(\cdot|x)}\left[\frac{p_\theta(x, Z)}{p_\theta(Z|x)}\right],$$

où $(x, z) \mapsto p_\theta(x, z)$ est la vraisemblance complète, $x$ sont les observations et $Z$ est la variable latente. Sous certaines hypothèses techniques, selon l'identité de Fisher, nous avons :

$$\nabla_\theta \log p_\theta(x) = \int \nabla_\theta \log p_\theta(x, z) p_\theta(z \mid x) \mathrm{d}z. \tag{5}$$

Cependant, dans la plupart des cas, la densité conditionnelle $z \mapsto p_\theta(z \mid x)$ est intractable et ne peut être échantillonnée directement. Les autoencodeurs variationnels introduisent un paramètre supplémentaire $\phi$ et une famille de distributions variationnelles $z \mapsto q_\phi(z \mid x)$ pour approcher la vraie distribution postérieure. Les paramètres sont estimés en maximisant la borne inférieure de la log-vraisemblance (ELBO) :

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x, Z)}{q_\phi(Z|x)}\right] =: \mathcal{L}_{\mathrm{ELBO}}(\theta, \phi; x).$$

L'algorithme IWAE [Burda et al., 2015] est une variante du VAE qui incorpore des poids d'importance pour obtenir une ELBO plus précise. L'objectif IWAE peut être écrit comme suit :

$$\mathcal{L}_k^{\mathrm{IWAE}}(\theta, \phi; x) = \mathbb{E}_{q_\phi^{\otimes k}(\cdot|x)} \left[ \log \frac{1}{k} \sum_{\ell=1}^{k} \frac{p_\theta(x, Z^{(\ell)})}{q_\phi(Z^{(\ell)}|x)} \right],$$

où $k$ correspond au nombre d'échantillons tirés sous la distribution variationnelle. L'estimateur du gradient de la ELBO dans IWAE est biaisé et le biais ainsi que l'erreur quadratique moyenne sont d'ordre $\mathcal{O}(1/k)$.

Nous menons nos expériences sur l'ensemble de données CIFAR-10 en utilisant Adagrad et RMSProp. Pour illustrer nos résultats, nous choisissons d'incorporer un biais d'ordre $\mathcal{O}(n^{-\alpha})$ à l'itération $n$. Comme le biais de l'estimateur du gradient dans IWAE est de l'ordre $\mathcal{O}(1/k)$, choisir un biais d'ordre $\mathcal{O}(n^{-\alpha})$ équivaut à utiliser $n^\alpha$ échantillons à l'itération $n$ pour estimer le gradient. Nous faisons varier $\alpha$ pour IWAE et nous affichons les quantités suivantes.

- Dans la Figure 1, la norme au carré du gradient $\|\nabla V(\theta_n)\|^2$ pour illustrer la vitesse de convergence.
- Dans la Figure 2, la vraisemblance négative en fonction des itérations.

La Figure 1 illustre nos résultats, tandis que les autres figures visent à confirmer le comportement de la perte de test avec différentes valeurs de $\alpha$. Toutes les figures sont tracées sur une échelle logarithmique pour une meilleure visualisation. Il est important de noter que toutes les figures sont par rapport aux époques, alors qu'ici, $n$ représente l'itération (nombre de mises à jour du gradient).

Nous pouvons clairement observer qu'une convergence rapide est atteinte dans chacun des cas lorsque $n$ est suffisamment grand. Il existe plusieurs explications possibles à cette convergence rapide. Par exemple, nous pourrions être en mesure d'améliorer la borne supérieure en obtenant une meilleure estimation du biais. Nos expériences montrent des résultats similaires pour Adagrad et RMSProp en termes de la vitesse de convergence, bien que RMSProp semble se comporter légèrement mieux.

Il est clair qu'avec un $\alpha$ plus grand, la convergence à la fois de la norme au carré du gradient et de la vraisemblance négative est plus rapide. Cependant, au-delà d'un certain seuil pour $\alpha$, nous observons que la vitesse de convergence ne change pas significativement. Comme choisir un $\alpha$ plus grand induit un coût computationnel supplémentaire, il est crucial de sélectionner une valeur appropriée de $\alpha$, qui atteint une convergence rapide sans être trop coûteuse en termes de temps de calcul.

# References

P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

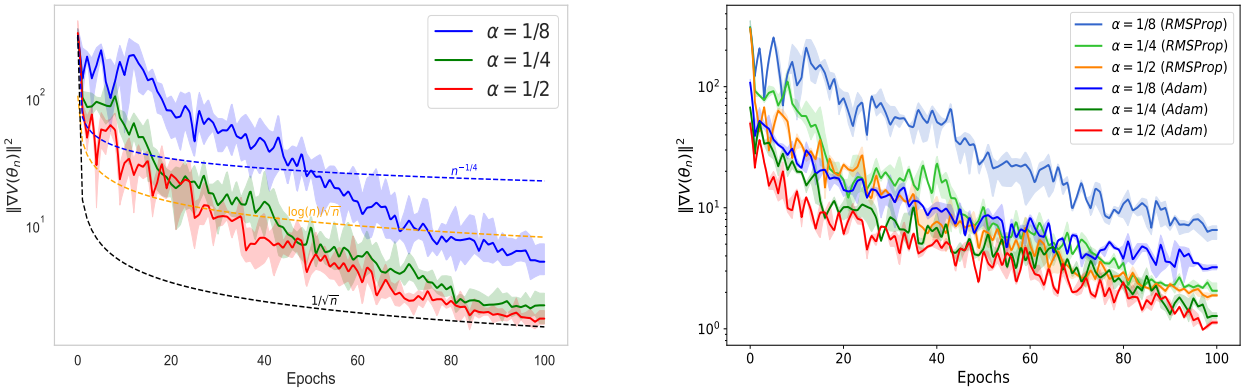



Figure 1: Valeur de $\|\nabla V(\theta_n)\|^2$ dans IWAE avec Adagrad (à gauche), RMSProp et Adam (à droite) pour différentes valeurs de $\alpha$. La courbe en pointillés correspond à la vitesse de convergence attendue $\mathcal{O}(n^{-1/4})$ pour $\alpha = 1/8$ et $\mathcal{O}(\log n/\sqrt{n})$ pour $\alpha = 1/4$ et pour $\alpha = 1/2$. Les lignes en gras représentent la moyenne sur 5 exécutions indépendantes.

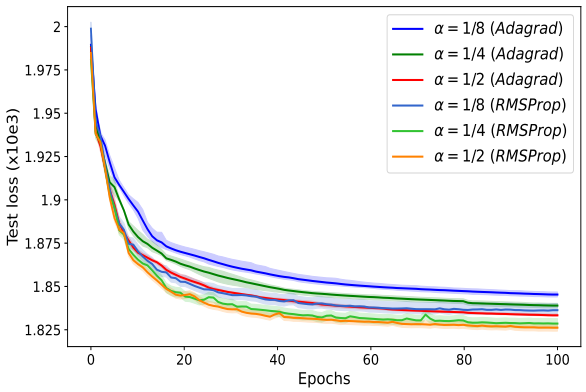



Figure 2: Vraisemblance négative sur l'ensemble de test de CIFAR-10 pour IWAE avec Adagrad and RMSProp en fonction des itérations pour différentes valeurs de $\alpha$.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

C. Boyer and A. Godichon-Baggioni. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972, 2023.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

G. Cardoso, Y. J. E. Idrissi, S. Le Corff, É. Moulines, and J. Olsson. State and parameter learning with PaRIS particle Gibbs. *arXiv preprint arXiv:2301.00900*, 2023.

T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg. Finite-time analysis of stochastic gradient descent under markov randomness. *arXiv preprint arXiv:2003.10973*, 2020.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

P. Gloaguen, S. Le Corff, and J. Olsson. A pseudo-marginal sequential Monte Carlo online smoothing algorithm. *Bernoulli*, 28(4):2606–2633, 2022.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

R. Leluc and F. Portier. Sgd with coordinate sampling: Theory and practice. *The Journal of Machine Learning Research*, 23(1):15470–15516, 2022.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24, 2011.

T. Sun, Y. Sun, and W. Yin. On Markov chain gradient descent. *Advances in Neural Information Processing Systems*, 31, 2018.

T. Tieleman, G. Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

F. Zou, L. Shen, Z. Jie, J. Sun, and W. Liu. Weighted adagrad with unified momentum. *arXiv preprint arXiv:1808.03408*, 2018.

# Régularisation implicite des réseaux de neurones profonds vers des EDO neuronales

Pierre Marion [1] & Yu-Han Wu [2] & Michael E. Sander [3] & Gérard Biau [4]

[1] *Institut de mathématiques, EPFL, Suisse,* pierre.marion@epfl.ch
[2] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France,* yu-han.wu@ens.psl.eu
[3] *École Normale Supérieure, CNRS, Département de Mathématiques et Applications, F-75005 Paris, France,* michael.sander@ens.fr
[4] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France,* gerard.biau@sorbonne-universite.fr

**Résumé.** Les réseaux neuronaux résiduels sont des modèles de pointe en apprentissage profond. Leur analogue à profondeur continue, les équations différentielles ordinaires (EDO) neuronales, sont également largement utilisées. Malgré leur succès, le lien entre les modèles discrets et continus manque encore d'une base mathématique solide. Dans cette contribution, nous faisons un pas dans cette direction en établissant une régularisation implicite des réseaux neuronaux résiduels profonds vers les EDO neuronales, pour des réseaux non linéaires entraînés avec un flot de gradient. Nous démontrons que si le réseau est initialisé comme une discrétisation d'une EDO neuronale, alors cette propriété est maintenue tout au long de l'entraînement. Nos résultats sont valides pour un temps d'entraînement fini, et également lorsque le temps d'entraînement tend vers l'infini à condition que le réseau satisfasse une condition de Polyak-Łojasiewicz. De plus, cette condition est vérifiée pour une famille de réseaux résiduels où les résidus sont des perceptrons à deux couches avec une surparamétrisation en largeur qui est seulement linéaire. Dans ce cas, nous montrons la convergence du flot de gradient vers un minimum global. Des expériences numériques illustrent nos résultats.

**Mots-clés.** Réseaux de neurones, apprentissage, flot de gradient, régularisation implicite

**Abstract.** Residual neural networks are state-of-the-art deep learning models. Their continuous-depth analog, neural ordinary differential equations (ODEs), are also widely used. Despite their success, the link between the discrete and continuous models still lacks a solid mathematical foundation. In this contribution, we take a step in this direction by establishing an implicit regularization of deep residual networks towards neural ODEs, for nonlinear networks trained with gradient flow. We prove that if the network is initialized as a discretization of a neural ODE, then such a discretization holds throughout training. Our results are valid for a finite training time, and also as the training time tends to infinity provided that the network satisfies a Polyak-Łojasiewicz condition. Importantly, this condition holds for a family of residual networks where the residuals are two-layer perceptrons with an overparameterization in width that is only linear, and implies the convergence of gradient flow to a global minimum. Numerical experiments illustrate our results.

**Keywords.** Neural networks, learning, gradient flow, implicit regularization

Nous nous intéressons aux propriétés des réseaux de neurones résiduels qui s'écrivent

$$\begin{aligned} h_0 &= Ax, \\ h_{k+1} &= h_k + \frac{1}{L} V_{k+1} \sigma(W_{k+1} h_k), \quad 0 \leqslant k \leqslant L-1, \\ F(x) &= B h_L, \end{aligned}$$

où la donnée est $x \in \mathbb{R}^d$, la matrice $A$ appartient à $\mathbb{R}^{q \times d}$, les états cachés $h_k$ sont dans $\mathbb{R}^q$, les matrices $V_{k+1}, W_{k+1}$ appartiennent respectivement à $\mathbb{R}^{q \times m}$ et $\mathbb{R}^{m \times q}$, et $B \in \mathbb{R}^{d' \times q}$. Nous considérons une initialisation dite régulière des paramètres $V_k$ et $W_k$, ce qui correspond à prendre les $V_k$ et $W_k$ comme des discrétisations de fonctions régulières (potentiellement aléatoires) $\mathcal{V} : [0,1] \to \mathbb{R}^{q \times m}$ et $\mathcal{W} : [0,1] \to \mathbb{R}^{m \times q}$, soit $V_k = \mathcal{V}(k/L)$ and $W_k = \mathcal{W}(k/L)$ pour $k \in \{1, \dots, L\}$. Cela inclut en particulier le cas où les $V_k$ et $W_k$ sont initialisées égales aux mêmes matrices $V$ et $W$ indépendamment de $k$. Dans ce cas, le réseau de neurones réalise à l'initialisation une discrétisation d'Euler de l'EDO

$$\begin{aligned} H(0) &= Ax, \\ \frac{dH}{ds}(s) &= \mathcal{V}(s) \sigma(\mathcal{W}(s) H(s)), \quad s \in [0,1], \\ F(x) &= BH(1). \end{aligned} \tag{1}$$

D'autres limites possibles à l'initialisation sont étudiées dans Marion *et al.* (2022).

Notre objectif dans ce travail (Marion *et al.*, 2024) est d'étudier le comportement du modèle dans le cas discuté ci-dessus, après entraînement. Nous montrons que les poids du réseau *entraîné* présentent toujours une structure de type EDO. Cette propriété était connue dans le cas des activations linéaires et dans un cadre plus restrictif (Sander *et al.*, 2022). Nous étendons ces résultats à un réseau résiduel non-linéaire assez général, qui se rapproche des réseaux utilisés en pratique. À cette fin, nous faisons l'hypothèse que le réseau est entraîné par flot de gradient, selon les équations d'évolution

$$\frac{\partial V_k}{\partial t}(t) = -L \frac{\partial \hat{\mathcal{R}}_n}{\partial V_k}(t), \quad \frac{\partial W_k}{\partial t}(t) = -L \frac{\partial \hat{\mathcal{R}}_n}{\partial W_k}(t), \quad t \geqslant 0,$$

où $\hat{\mathcal{R}}_n$ désigne un risque empirique. Notons en particulier que la variable temporelle $t$ de l'EDO qui décrit l'évolution des poids n'est pas la même que la variable $s$ de l'EDO neuronale (1) qui décrit la limite en large profondeur.

Notre première contribution est de prouver que la convergence (lorsque $L$ tend vers l'infini) du réseau résiduel vers une EDO neuronale est également valide après entraînement. Cette convergence est valide pour tout temps d'entraînement fini $t \in [0, T]$. Cette propriété est en fait valide dans un cadre beaucoup plus général, dès lors que le réseau de neurones s'écrit comme

$$h_{k+1}^L = h_k^L + \frac{1}{L} f(h_k^L, Z_{k+1}^L), \quad k \in \{0, \dots, L-1\},$$

où $f : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}^q$ est une fonction $\mathcal{C}^2$ telle que $f(0, \cdot) \equiv 0$ and $f(\cdot, z)$ est uniformément Lipschitz pour $z$ dans tout compact.

Néanmoins, la convergence de l'algorithme d'optimisation lorsque $T$ tend vers l'infini n'est pas garantie sans hypothèse supplémentaire, du fait de la non-convexité du problème d'optimisation. Nous prouvons cette convergence grâce à une condition de type Polyak-Łojasiewicz (PL), un outil majeur dans l'analyse des algorithmes d'optimisation pour les réseaux de neurones (Liu *et al.*, 2022). La condition PL implique la convergence du flot de gradient vers un minimum global. Notre seconde contribution est de prouver que cette condition est vérifiée lorsque la largeur $q$ des couches cachées est plus grande qu'une constante fois la taille de l'échantillon $n$. Cette condition de surparamétrisation est meilleure que les résultats de la littérature, qui requiert soit une surparamétrisation polynomiale, soit des conditions plus restrictives sur les données. D'autres hypothèses légères sont nécessaires, en particulier sur la forme exacte de l'initialisation. Nous obtenons alors la convergence en grande profondeur et en grand temps d'entraînement, c'est-à-dire l'existence de fonctions Lipschitz $\mathcal{V}_\infty$ et $\mathcal{W}_\infty$ telles que le réseau de neurones entraîné converge lorsque $L$ *et* $T$ tendent vers l'infini vers l'EDO

$$\frac{dH}{ds}(s) = \mathcal{V}_\infty(s)\sigma(\mathcal{W}_\infty(s)H(s)), \quad s \in [0,1]. \tag{2}$$

De plus, l'erreur d'entraînement de l'EDO limite est égale à zéro. Cette analyse représente une première étape dans la compréhension de la régularisation implicite du flot de gradient pour les réseaux résiduels, c'est-à-dire la caractérisation des propriétés du réseau entraîné parmi tous les minimiseurs du risque empirique.

Nos résultats théoriques sont complétés par des illustrations expérimentales, qui montrent en particulier qu'il est possible d'apprendre avec un réseau initialisé comme décrit au début de ce document. Nous obtenons ainsi une performance de l'ordre de 80% sur CIFAR-10, et les poids après entraînement réalisent bien une discrétisation d'une fonction lisse (Figure 1 à gauche), c'est-à-dire que le réseau entraîné discrétise bien une EDO.

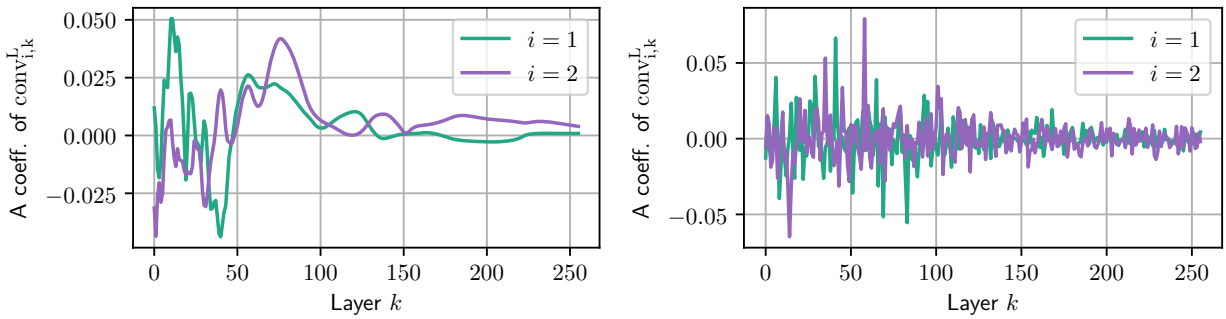



FIGURE 1 – Un coefficient aléatoire des matrices de poids suivi le long de la profondeur du réseau, après entraînement. **Gauche :** L'initialisation lisse des poids conduit à des poids qui discrétisent une fonction lisse après entraînement. **Droite :** En initialisant les poids de manière i.i.d., nous obtenons des poids non lisses après entraînement.

# Références

C. LIU, L. ZHU et M. BELKIN : Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*,

59:85–116, 2022. Special Issue on Harmonic Analysis and Machine Learning.

P. Marion, A. Fermanian, G. Biau et J.-P. Vert : Scaling ResNets in the large-depth regime. *arXiv :2206.06929*, 2022.

P. Marion, Y.-H. Wu, M.E. Sander et G. Biau : Implicit regularization of deep residual networks towards neural ODEs. *In International Conference on Learning Representations*, 2024.

M.E. Sander, P. Ablin et G. Peyré : Do residual neural networks discretize neural ordinary differential equations ? *In* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, éditeurs : *Advances in Neural Information Processing Systems*, volume 35, pages 36520–36532. Curran Associates, Inc., 2022.

# Régression sur variables entachées d'erreurs à l'aide de réseaux de neurones bayésiens : présentation d'une approche motivée par la datation carbone 14

Destin Ashuza Cirumanga [1], Anne Philippe [2] & Guillaume Guérin [3]

[1] *CNRS & Laboratoire de Mathématiques Jean Leray, Nantes Université, France.*
*destin.ashuzacirumanga@univ-nantes.fr*
[2] *Laboratoire de Mathématiques Jean Leray, Nantes Université, France.*
*anne.philippe@univ-nantes.fr*
[3] *CNRS & Géosciences Rennes, Université Rennes 1, France.*
*guillaume.guerin@univ-rennes1.fr*

**Résumé.** L'estimation d'une fonction de régression par un modèle paramétrique en présence d'une variable explicative bruitée conduit à une estimation biaisée et non consistante si on ne prend pas en compte l'incertitude en entrée du modèle. Des solutions existent pour la prise en compte de cette incertitude dans le cadre des modèles linéaires et non linéaires classiques. Cependant, elles ne sont pas adaptées aux réseaux de neurones. En utilisant l'approximation variationnelle gaussienne combinée à une modélisation jointe de la variable explicative entachée d'erreur et de la variable dépendante, nous proposons une approche permettant de prendre en compte l'incertitude en entrée du réseau tout en préservant une simplicité d'entraînement similaire à celle du cas standard d'un apprentissage sur des entrées non bruitées. De plus cet apprentissage peut se faire en mode distribué. Nous validons notre approche sur des données simulées ayant des caractéristiques similaires aux données réelles de datation par le carbone 14 en archéologie qui ont motivé ce travail.

**Mots-clés.** Réseaux de neurones, statistique bayésienne, inférence variationnelle, variables entachées d'erreurs.

**Abstract.** Estimation of a regression function by a parametric model in the presence of a noisy input variable leads to a biased and inconsistent estimate if the uncertainty in the model input is not taken into account. Solutions exist for taking this uncertainty into account in the context of classical linear and non-linear models. However, they are not suitable for neural networks. By using the Gaussian variational approximation combined with joint modeling of the error-prone input variable and the dependent variable, we propose an approach that takes into account the uncertainty in the network input while preserving a training simplicity similar to that of the standard case of learning on noiseless inputs. What's more, this training can be carried out in distributed mode. We validate our approach on simulated data with characteristics similar to the real archaeological carbon-14 dating data that motivated this work.

**Keywords.** Neural networks, Bayesian statistics, variational inference, errors-in-variables models.

# 1 Introduction à la problématique des variables explicatives entachées d'erreurs en régression

En apprentissage statistique, quand on veut estimer un modèle de régression, il est fréquent de considérer que les variables explicatives sont déterministes et que seule la variable à expliquer est bruitée. Cependant, dans certaines situations, les variables explicatives peuvent être entachées d'erreurs. La problématique de l'estimation de la courbe de calibration des âges carbone 14 traitée dans Heaton et al. (2020) en est un exemple. En effet, la calibration des âges carbone 14 nécessite l'estimation d'une fonction de régression qui prend en entrée la date (âge) d'un objet donné et donne en sortie l'âge carbone 14 correspondant. Mais les dates utilisées sont entachées d'erreurs pour certaines périodes de temps.

La présence d'incertitude dans les variables explicatives conduit aux modèles dits de régression sur variables entachées d'erreurs. Ces modèles ont été largement étudiés dans la littérature dans le cadre de la régression linéaire. Un aperçu de l'état de l'art sur la question peut être trouvé dans Carroll et al. (2006). Un constat est que la non prise en compte de l'erreur dans une variable explicative conduit à une estimation biaisée des paramètres du modèle. Ce phénomène est connu sous le nom de biais d'atténuation.

De nombreuses méthodes pour la prise en compte de l'erreur dans une variable explicative ont été proposées pour les modèles linéaires et les modèles linéaires généralisés. C'est le cas par exemple des méthodes de régression - calibration, simulation - extrapolation, variable instrumentale ou certaines fonctions de score. Comme souligné dans Carroll et al. (2006), ces méthodes ont des performances faibles pour des modèles fortement non linéaires. Pour les modèles plus complexes ou fortement non linéaires, des alternatives existent parmi lesquelles des approches basées sur les méthodes de vraisemblance ou les modèles bayésiens hiérarchiques. Dans Heaton et al. (2020) par exemple, une approche bayésienne hiérarchique couplée aux splines est utilisée pour estimer la courbe de calibration des âges carbone 14. Notre objectif est de proposer une estimation de cette courbe par des réseaux de neurones bayésiens.

La problématique de la prise en compte de l'incertitude dans une variable explicative pour les réseaux de neurones a été peu étudiée. Martin and Elster (2023) propose une approche basée sur l'algorithme de Monte Carlo dropout. Nous proposons dans ce papier une alternative basée sur l'approximation variationnelle gaussienne. La suite de ce papier se subdivise en trois points : dans la section 2 nous formalisons le problème et dérivons la fonction de perte à optimiser pour estimer le modèle ; sur base de cette fonction de perte, dans la section 3 nous proposons deux stratégies d'apprentissage permettant de prendre en compte l'incertitude en entrée du réseau tout en conservant une grande simplicité dans l'entraînement du réseau ; et enfin dans la section 4 nous présentons les résultats obtenus par notre approche sur des données simulées.

## 2 Modélisation et estimation de la fonction de régression

Nous disposons des données d'apprentissage $\mathcal{D} = \{(t_i, \sigma_i, M_i, s_i), i = 1, \cdots, n\}$ avec
— $d_i$ : la vraie valeur de la variable explicative inconnue,
— $t_i$ : la variable explicative $d_i$ entachée d'erreur,
— $\sigma_i$ : l'écart-type de l'erreur associée à $t_i$,
— $M_i$ : la variable à expliquer bruitée, et
— $s_i$ : l'écart-type du bruit associé à $M_i$.
Les variables observées $t_i$ et $M_i$ sont distribuées suivant des lois normales de telle sorte que le modèle qui en découle s'écrive comme suit :

$$\begin{cases} M_i = m_i + s_i \cdot \varepsilon_i \\ m_i = g(d_i, \theta) \\ t_i = d_i + \sigma_i \cdot u_i \end{cases} \tag{1}$$

avec $\varepsilon_i \sim \mathcal{N}(0,1)$, $u_i \sim \mathcal{N}(0,1)$, $g$ la fonction de régression que l'on souhaite estimer par un réseau de neurones bayésien de paramètres (les poids et biais du réseau) $\theta$ dont la loi *a priori* $\pi(\theta)$ est une gaussienne multivariée de composantes toutes indépendantes, centrées et réduites. En l'absence des connaissances particulières, nous choisissons une loi *a priori* non informative pour $d_i$ : la loi de Laplace i.e $\pi(d_i) \propto \mathbb{1}_{\mathbb{R}}(d_i)$.

Plus précisément, si le réseau de neurones est composé de $L$ couches, la fonction de régression $g$ s'écrit alors comme suit :

$$g(d_i, \theta) = \Phi_L \circ \Phi_{L-1} \circ \cdots \circ \Phi_1(X_0) \tag{2}$$

avec $\theta = \{(\omega_j, b_j)_{1 \leq j \leq L}\}$, l'entrée $X_0 = d_i$ et pour toute couche $j \in \{1, \cdots, L\}$, le vecteur des sorties $X_j$ est donné par $X_j = \Phi_j(X_{j-1}) = \phi_j(\omega_j X_{j-1} + b_j)$, $\omega_j$ et $b_j$ étant respectivement la matrice des poids et le vecteur des biais de neurones de cette couche, et $\phi_j$ sa fonction d'activation. Pour tout $x \in \mathbb{R}$, on a $\phi_L(x) = x$ et $\phi_j(x) = \max(0, x)$ pour chaque $j \in \{1, \cdots, L-1\}$ (la fonction $ReLU$). Dans la notation $\phi_j(\omega_j X_{j-1} + b_j)$, la fonction d'activation $\phi_j$ est évaluée composante par composante. Ces spécifications relatives à la fonction $g$ définissent un perceptron multicouche pour notre problème. Plus de détails sur les perceptrons et d'autres architectures de réseaux de neurones peuvent être trouvés dans James et al. (2021).

Pour spécifier intégralement notre modèle bayésien, il reste à déterminer sa vraisemblance. Comme présentées dans Carroll et al. (2006), les méthodes de vraisemblance pour les modèles de régression sur variables entachées d'erreurs sont basées sur le calcul de la densité jointe de la variable à expliquer et de la variable explicative entachée d'erreur. Pour tout $i \in \{1, \cdots, n\}$, la densité $p(M_i, t_i | \theta, s_i, \sigma_i)$ de $(M_i, t_i)$ s'écrit :

$$\begin{aligned} p(M_i, t_i | \theta, s_i, \sigma_i) &= \int p(M_i, t_i, d_i | \theta, s_i, \sigma_i)\, d(d_i) \\ &= \int p(M_i | t_i, d_i, \theta, s_i, \sigma_i)\, p(t_i, d_i | \theta, s_i, \sigma_i)\, d(d_i) \\ &= \int p(M_i | d_i, \theta, s_i)\, p(t_i | d_i, \sigma_i)\, \pi(d_i)\, d(d_i) \end{aligned} \tag{3}$$

Le passage de la deuxième égalité à la troisième s'explique par le fait que la loi conditionnelle de $M_i$ sachant $(d_i, \theta, s_i)$ ne dépend pas de $(t_i, \sigma_i)$ et celle de $t_i$ sachant $(d_i, \sigma_i)$ ne dépend pas de $(\theta, s_i)$. Par la formule de Bayes, nous avons également :

$$\pi(d_i|t_i, \sigma_i) \propto p(t_i|d_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d_i - t_i}{\sigma_i}\right)^2\right] \tag{4}$$

Il en découle que la loi *a posteriori* $\pi(d_i|t_i, \sigma_i)$ est la loi gaussienne $N(t_i, \sigma_i^2)$. L'égalité (3) devient alors

$$\begin{aligned} p(M_i, t_i|\theta, s_i, \sigma_i) &= \int p(M_i|d_i, \theta, s_i)\, \pi(d_i|t_i, \sigma_i)\, d(d_i) \\ &= \mathbb{E}_{\pi(d_i|t_i,\sigma_i)}\left[p(M_i|d_i, \theta, s_i)\right] \end{aligned} \tag{5}$$

Un estimateur de Monte Carlo de cette densité jointe est donnée alors par

$$\begin{aligned} \widehat{p}_K(M_i, t_i|\theta, s_i, \sigma_i) &= \frac{1}{K}\sum_{k=1}^{K} p(M_i|\theta, d_i^k, s_i) \\ &= \frac{1}{s_i\sqrt{2\pi}}\frac{1}{K}\sum_{k=1}^{K} exp\left\{-\frac{1}{2}\frac{\left(M_i - g(d_i^k, \theta)\right)^2}{s_i^2}\right\} \end{aligned} \tag{6}$$

avec $\left(d_i^k\right)$, $k = \{1, \cdots, K\}$ i.i.d de loi $N(t_i, \sigma_i^2)$.

Finalement, en utilisant (5) et (6) et en notant $\underline{M} = (M_1, \cdots, M_n)$, $\underline{s} = (s_1, \cdots, s_n)$, $\underline{t} = (t_1, \cdots, t_n)$, $\underline{\sigma} = (\sigma_1, \cdots, \sigma_n)$ et $\underline{d} = (d_1, ..., d_n)$ les vecteurs contenant les différentes données et variables, la vraisemblance du modèle est alors donnée par

$$\begin{aligned} \mathcal{L}(\theta) &= p(\underline{M}, \underline{t}|\theta, \underline{s}, \underline{\sigma}) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n}\frac{1}{s_i}\mathbb{E}_{\pi(d_i|t_i,\sigma_i)}\left[\exp\left\{-\frac{1}{2}\frac{\left(M_i - g(d_i, \theta)\right)^2}{s_i^2}\right\}\right] \end{aligned} \tag{7}$$

et son estimateur par, pour tout entier $K \geq 1$,

$$\widehat{\mathcal{L}}_K(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{K^n}\prod_{i=1}^{n}\frac{1}{s_i}\sum_{k=1}^{K} exp\left\{-\frac{1}{2}\frac{\left(M_i - g(d_i^k, \theta)\right)^2}{s_i^2}\right\} \tag{8}$$

avec $\left(d_i^k\right)$, $k = \{1, \cdots, K\}$ et $i = \{1, \cdots, n\}$ indépendantes et pour tout $i \in \{1, \cdots, n\}$, $\left(d_i^k\right)_{1\leq k\leq K}$ i.i.d de loi $N(t_i, \sigma_i^2)$.

Dans le cadre bayésien, l'inférence et la prédiction du modèle décrit par le système d'équations (1) sont basées sur la loi *a posteriori* $\pi(\theta|\mathcal{D})$ sur les paramètres du réseau. Cette loi s'obtient par la formule de Bayes :

$$\pi(\theta|\mathcal{D}) = \frac{\mathcal{L}(\theta)\cdot\pi(\theta)}{m(\underline{M}, \underline{t}|\underline{s}, \underline{\sigma})} \tag{9}$$

avec $m(\underline{M}, \underline{t}|\underline{s}, \underline{\sigma}) = \int \mathcal{L}(\theta) \cdot \pi(\theta) d\theta$ . Cette constante de normalisation étant en général difficile à calculer, les méthodes de Monte Carlo par Chaînes de Markov (MCMC) sont alors utilisées dans les cas de modèles bayésiens classiques afin de générer un échantillon suivant la loi *a posteriori*. Cependant, les réseaux de neurones ont souvent un nombre de paramètres très élevé, ce qui rend l'utilisation des algorithmes MCMC impossible dans un laps de temps raisonnable. La solution est alors de recourir à l'inférence variationnelle :

- son principe : remplacer la loi *a posteriori* $\pi(\theta|\mathcal{D})$ par une loi $\pi(\theta|\psi)$ « plus simple à manipuler » appelée distribution variationnelle,
- le but sera alors de trouver le paramètre $\psi$ tel que la dissimilarité entre ces deux lois soit minimale,
- pour ce faire, on minimise la divergence de Kullback Leibler, notée $KL(\cdot \,||\, \cdot)$, entre les deux distributions :

$$\begin{aligned} KL\big(\pi(\theta|\psi) \,||\, \pi(\theta|\mathcal{D}\big) &= \int \pi(\theta|\psi) \log \frac{\pi(\theta|\psi)}{\pi(\theta|\mathcal{D})} d\theta \\ &= \mathbb{E}_{\pi(\theta|\psi)} \left[\log \frac{\pi(\theta|\psi)}{\pi(\theta|\mathcal{D})}\right] \\ &= KL\big(\pi(\theta|\psi) \,||\, \pi(\theta)\big) - \mathbb{E}_{\pi(\theta|\psi)} \left[\log\left(\mathcal{L}(\theta)\right)\right] + \log\left(m(\underline{M}, \underline{t}|\underline{s}, \underline{\sigma})\right) \end{aligned} \tag{10}$$

L'équation (10) montre alors que minimiser la divergence de Kullback Leibler entre la loi *a posteriori* et la loi variationnelle revient tout simplement à minimiser la fonction de perte $\ell(\psi)$ donnée par

$$\ell(\psi) = KL\big(\pi(\theta|\psi) \,||\, \pi(\theta)\big) - \mathbb{E}_{\pi(\theta|\psi)} \left[\log\left(\mathcal{L}(\theta)\right)\right] \tag{11}$$

L'opposé de cette fonction de perte est connu dans la littérature sous l'acronyme **ELBO** pour *Evidence Lower Bound*, voir par exemple dans Jaakkola and Jordan (2000). Le livre Bishop (2006) contient un bon chapitre introductif sur l'approche variationnelle et Blei et al. (2017) en donne une revue détaillée.

Nous pouvons alors recourir à une double approximation par Monte Carlo pour calculer un estimateur $\widetilde{\ell}_{K,Q}(\psi)$ de la fonction de perte donnée par (11) : on remplace d'abord la vraisemblance $\mathcal{L}(\theta)$ par son estimateur $\widehat{\mathcal{L}}_K(\theta)$ donné par (8), ensuite on estime $\mathbb{E}_{\pi(\theta|\psi)} \left[\log\left(\widehat{\mathcal{L}}_K(\theta)\right)\right]$ en simulant un $Q$-échantillon $(\theta_q)$, $q = \{1, \cdots, Q\}$ suivant la loi variationnelle $\pi(\theta|\psi)$. On obtient alors

$$\widetilde{\ell}_{K,Q}(\psi) = C_n + KL\big(\pi(\theta|\psi) \,||\, \pi(\theta)\big) - \sum_{i=1}^{n} \frac{1}{Q} \sum_{q=1}^{Q} \log\left(\frac{1}{K} \sum_{k=1}^{K} exp\left\{-\frac{1}{2} \frac{\left(M_i - g(d_i^k, \theta_q)\right)^2}{s_i^2}\right\}\right) \tag{12}$$

avec $C_n = n \log(\sqrt{2\pi}) + \sum_{i=1}^{n} \log(s_i)$ une constante.

Et en appliquant l'inégalité de Jensen à la fonction $-\log$ qui est convexe, on obtient la majoration

$$\widetilde{\ell}_{K,Q}(\psi) \leq \widehat{\ell}_{K,Q}(\psi) \tag{13}$$

dans laquelle l'expression analytique de la fonction majorante $\widehat{\ell}_{K,Q}(\psi)$ est donnée par

$$\widehat{\ell}_{K,Q}(\psi) = C_n + KL\big(\,\pi(\theta|\psi)\,||\,\pi(\theta)\,\big) + \frac{1}{2}\sum_{i=1}^{n}\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{K}\sum_{k=1}^{K}\frac{\big(M_i - g(d_i^k,\theta_q)\big)^2}{s_i^2} \tag{14}$$

C'est cette dernière fonction $\widehat{\ell}_{K,Q}$ que nous allons utiliser comme fonction de perte. En effet, l'estimateur $\widetilde{\ell}_{K,Q}$ donné par (12) est certes calculable mais nécessiterait une bouche d'apprentissage conçue sur mesure car ce n'est pas une fonction standard qu'on retrouve dans les librairies classiques de deep learning. Son évaluation peut aussi avoir un coût de calcul non négligeable selon l'efficacité de son implémentation. A l'opposé, avec le choix de $Q = 1$, utiliser $\widehat{\ell}_{K,Q}$ comme fonction de perte lors de l'entraînement du réseau est équivalent à minimiser l'erreur quadratique moyenne (MSE) pondérée d'un réseau avec variable explicative non entachée d'erreur et entrainé sur $n \cdot K$ données $\mathcal{D}_{nK} = \big\{(d_i^k, M_i, s_i), i = 1, \cdots, n \texttt{ et } k = 1, \cdots, K\big\}$ où les $d_i^k$ ont été générées en amont suivant les lois $N(t_i, \sigma_i^2)$. Et à cette MSE pondérée est ajouté un terme de régularisation donné par la divergence de Kullback Leibler entre la loi variationnelle et la loi *a priori*. De plus, l'entraînement peut être séquentiel ou distribué ; ce qui donne deux stratégies d'apprentissage présentées dans la section 3 ci-dessous. Cependant, minimiser la fonction de remplacement $\widehat{\ell}_{K,Q}$ ne garantit pas forcément l'obtention d'une solution optimale pour la vraie fonction de perte $\widetilde{\ell}_{K,Q}$ qu'elle majore. Nous avons donc testé l'utilisation de cette approche sur des données simulées pour évaluer sa capacité à estimer la vraie fonction de régression. Les résultats obtenus sont l'objet de la section 4.

La divergence entre la loi variationnelle et la loi *a priori* $KL\big(\,\pi(\theta|\psi)\,||\,\pi(\theta)\,\big)$ qui apparaît comme terme de régularisation de la fonction de perte dans (11), (12), et (14) peut être approchée numériquement par Monte Carlo comme l'espérance, sous la loi variationnelle, de $\log\dfrac{\pi(\theta|\psi)}{\pi(\theta)}$. Dans ce travail, nous avons choisi des lois variationnelles gaussiennes indépendantes sur les poids et biais du réseau de neurones. Dans ce cas gaussien pour la loi variationnelle et la loi *a priori* sur les paramètres, on peut même donner une expression analytique explicite de la divergence de Kullback Leibler entre les deux lois. Si on réécrit $\theta = (\theta_1, \cdots, \theta_S)$ et $\psi = (\psi_1, \cdots, \psi_S) = ((\psi_{1,1}, \psi_{1,2}), \cdots, (\psi_{S,1}, \psi_{S,2}))$, où $S$ représente le nombre de paramètres du réseau, on peut montrer que

$$KL\big(\,\pi(\theta|\psi)\,||\,\pi(\theta)\,\big) = -\frac{S}{2} + \sum_{s=1}^{S}\left[\frac{\psi_{s,1}^2}{2} + \frac{\psi_{s,2}^2}{2} - \log\left(\psi_{s,2}\right)\right] \tag{15}$$

Maintenant que tous les termes de la fonction de perte sont facilement calculables, l'entraînement du réseau peut se faire comme d'habitude par rétro-propagation des gradients en utilisant par exemple le très populaire astuce de ré-paramétrisation utile dans le contexte de l'approche variationnelle et conduisant à l'algorithme Bayes-by-Backprop introduite par Blundell et al. (2015). Quant au choix de $Q = 1$ mentionné ci-haut, Kingma and Welling (2014) expliquent qu'il est valable à condition que la taille des mini batchs utilisés lors de l'entraînement du réseau soit grande, par exemple 100.

# 3 Stratégies d'apprentissage

Nous venons de voir que l'utilisation de la fonction de perte définie par (14) permet d'entrainer le réseau sur les données $\mathcal{D}_{nK} = \left\{(d_i^k, M_i, s_i), i = 1, \cdots, n \texttt{ et } k = 1, \cdots, K\right\}$ où les $d_i^k$ ont été générées en amont suivant les lois $N(t_i, \sigma_i^2)$. Nous avons alors deux possibilités pour l'entraînement du réseau : un apprentissage séquentiel ou un apprentissage distribué.

## Apprentissage séquentiel

En utilisant les différentes lois *a priori* et *a posteriori* spécifiées précédemment ainsi que la distribution variationnelle gaussienne, l'apprentissage séquentiel se fait en deux étapes simples :

— étape 1 : choisir une valeur de $K$ et générer les données $\left\{(d_i^k, M_i, s_i)\right\}$, $i = 1, \cdots, n$ et $k = 1, \cdots, K$, où les $d_i^k \sim N(t_i, \sigma_i^2)$.

— étape 2 : entraîner un réseau bayésien de manière habituelle avec ces données en choisissant comme fonction de perte la MSE pondérée régularisée par la divergence de Kullback Leibler. Ceci se fait très simplement en utilisant les librairies de deep learning classiques, par exemple Keras et Tensorflow-Probability.

## Apprentissage distribué

Le fait de passer d'un jeu des données de taille à $n$ à un jeu des données de taille d'ordre $n \cdot K$ peut-être couteux en temps de calcul, surtout si $n$ et $K$ sont grands. L'idée est alors de subdiviser le jeu des données $\left\{(d_i^k, M_i, s_i)\right\}$ en $K$ jeux de données pour entraîner $K$ réseaux de neurones distincts et indépendants mais possédant la même architecture (nombre de couches, neurones par couche et fonctions d'activation). Le réseau 1 utilise les données $\{(d_i^1, M_i, s_i)\}$, le réseau 2 les données $\{(d_i^2, M_i, s_i)\}$, ainsi de suite jusqu'au réseau $K$ qui va utiliser les données $\left\{(d_i^K, M_i, s_i)\right\}$ avec $i = 1, \cdots, n$ pour chaque jeu de données. L'entraînement de chaque réseau se fait alors suivant les étapes de l'apprentissage séquentiel avec une petite modification sur la loi à priori $\pi(\theta)$ : au lieu de prendre des gaussiennes centrées réduites, on prend des lois gaussiennes centrées mais de variance $K^{-1}$ pour permettre la reconstruction de la loi *a posteriori* initiale, et donc de la loi variationnelle dans notre cas. Ces réseaux étant indépendants, ils peuvent être entrainés en parallèle selon les ressources de calcul disponibles.

Une fois les réseaux entrainés, il faut recombiner leurs lois variationnelles pour former la loi variationnelle du réseau unique qui aurait pu être entraîné sur toutes ces données en mode apprentissage séquentiel. Alors pour tout paramètre $\theta_s$ du réseau unique, sa distribution variationnelle reconstruite à partir de celles de $K$ réseaux est la loi normale de moyenne $\psi_{s,1}$ et de variance $\psi_{s,2}^2$ avec

$$\psi_{s,2}^2 = \left(\sum_{k=1}^{K} \frac{1}{\psi_{s,2,k}^2}\right)^{-1} \quad \texttt{et} \quad \psi_{s,1} = \psi_{s,2}^2 \sum_{k=1}^{K} \frac{\psi_{s,1,k}}{\psi_{s,2,k}^2}$$

où pour tout $1 \leq k \leq K$, $\psi_{s,1,k}$ et $\psi_{s,2,k}^2$ représentent respectivement la moyenne et la variance

de la loi variationnelle gaussienne du paramètre $\theta_s$ dans le $k$ ième réseau de neurone. Cette méthode est par exemple utilisée par Huang and Gelman (2005) pour reconstruire la loi *a posteriori* par approximation gaussienne lors d'un apprentissage distribué.

Pour améliorer les performances du réseau en termes de temps de calcul et réduire aussi le nombre de paramètres qui croit très vite avec la taille du réseau dans le cas bayésien, une méthode pratique mentionnée dans Jospin et al. (2022) consiste à ne faire du bayésien que sur les dernières couches du réseau. Nous avons appliqué cette méthode dans Ashuza Cirumanga et al. (2023) pour proposer une méthode de calibration des âges carbone 14. Si on veut appliquer l'apprentissage distribué dans ce cas de figure, on peut d'abord entraîner un réseau non bayésien une fois sur toutes les données. Ensuite il suffit de figer les paramètres de toutes les couches non bayésiennes et faire de l'apprentissage distribué pour les couches bayésiennes.

# 4 Expérimentations numériques et conclusion

Nous appliquons l'approche proposée sur des données simulées. Le but est de vérifier si l'on arrive à estimer la vraie fonction de régression au cœur du processus de génération des données. Toutefois, nous veillons à ce que les données simulées aient des caractéristiques similaires (variation de la fonction, structure et taille des erreurs par rapport aux variables explicative et à expliquer) aux données de datation carbone 14 afin de pouvoir appliquer la même approche à l'estimation de courbe de calibration des âges carbone 14 pour les périodes où les dates sont entachées d'erreurs ; ce qui complétera le travail présenté dans Ashuza Cirumanga et al. (2023) qui portait sur l'estimation de cette courbe pour les périodes de temps où la chronologie est connue de manière absolue.

La fonction $g$ utilisée pour générer les données est alors donnée par :

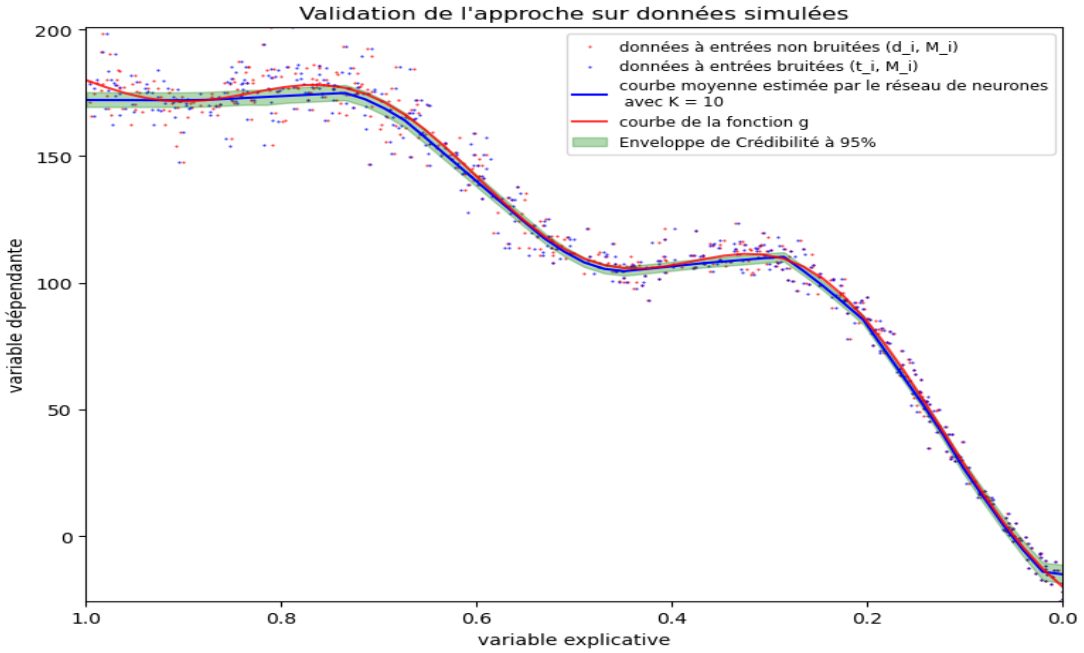
$$\forall x \in [0,1], g(x) = \frac{b-a}{2}\left[x + \frac{1}{3}\left(\sin^2(2\pi x) + \cos\left((1-2x)\frac{\pi}{2}\right) + x^3\cos^2(2\pi x)\right)\right] + a$$

Les données d'apprentissage $\mathcal{D} = \{(t_i, \sigma_i, M_i, s_i), i = 1, \cdots, n\}$ sont alors obtenues comme suit : pour tout $i \in \{1, \cdots, n\}$ :

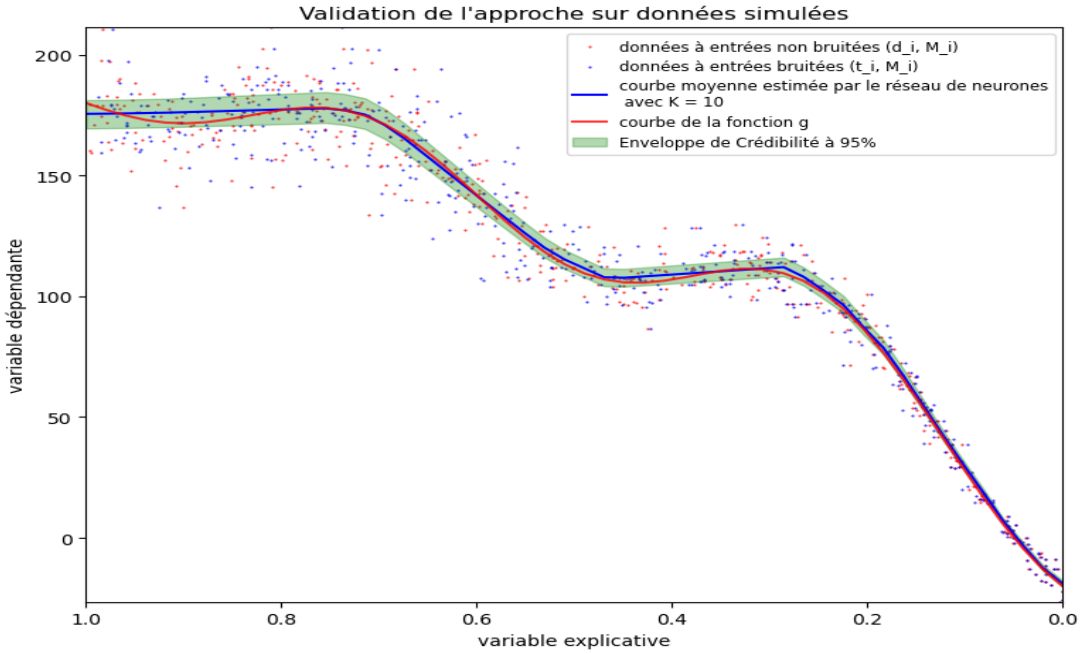
- on tire la vraie valeur $d_i$ de la variable explicative suivant une loi uniforme sur $[0,1]$,
- $\sigma_i = (\alpha + \beta v_i)|d_i|$ avec $v_i$ tiré uniformément sur $[0,1]$,
- $t_i = d_i + \sigma_i u_i$ avec $u_i \sim N(0,1)$,
- $s_i = (\lambda + \kappa w_i)|g(d_i)|$ avec $w_i$ tiré uniformément sur $[0,1]$, et
- $M_i = g(d_i) + s_i\,\varepsilon_i$ avec $\varepsilon_i \sim N(0,1)$.

Pour générer les données, nous avons fixé $a = -20$ et $b = 280$. $\alpha$, $\beta$, $\lambda$ et $\kappa$ sont des réels choisis dans $[0,1]$ de telle sorte que les erreurs sur les variables explicative et à expliquer soient de taille voulue (en pourcentage par rapport à la variable concernée).

Sur la figure 1, le graphique (1a) montre le résultat obtenu pour l'estimation de la fonction $g$ quand l'erreur commise sur la variable explicative est de l'ordre de 0.3% à 2.3% et celle sur la variable à expliquer de 2% à 7%. Ces ordres d'erreur correspondent à ceux qui sont observés dans les données de datation des âges carbone 14 au delà des 12 derniers millénaires. Le graphique (1b) montre la même chose pour des ordres d'erreur un peu plus grand : 3% à 5%




(a) $\alpha = 0.003$, $\beta = 0.02$, $\lambda = 0.02$, $\kappa = 0.05$




(b) $\alpha = 0.03$, $\beta = 0.02$, $\lambda = \kappa = 0.05$

FIGURE 1 – Estimation de $g$ avec un réseau de neurones hybride à trois couches cachées contenant respectivement 20, 260 et 240 neurones pour les trois couches successives. Seule la couche de sortie est bayésienne. $n = 1500$ observations utilisées pour l'apprentissage.

pour la variable explicative et 5% à 10% pour la variable à expliquer. Dans les deux cas, la qualité d'estimation de la courbe est satisfaisante et l'approche peut donc être utilisée pour

estimer la courbe de calibration des âges carbone 14. Le graphique (1b) illustre bien aussi l'impact des erreurs en entrée sur la largeur de l'enveloppe de crédibilité obtenue autour de la courbe.

# Références


Ashuza Cirumanga, D., Philippe, A., and Guérin, G. (2023). Approche bayésienne et réseaux de neurones appliqués à la calibration des âges carbone 14. In *communication lors des 54es Journées de la Statitistique de la SFdS*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference : A review for statisticians. *Journal of the American statistical Association*, 112(518) :859–877.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models : a modern perspective*, volume 2. Chapman and Hall/CRC.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. (2017). Tensorflow distributions. *arXiv preprint arXiv :1711.10604*.

Heaton, T. J., Blaauw, M., Blackwell, P. G., Ramsey, C. B., Reimer, P. J., and Scott, E. M. (2020). The intcal20 approach to radiocarbon calibration curve construction : a new methodology using bayesian splines and errors-in-variables. *Radiocarbon*, 62(4) :821–863.

Huang, Z. and Gelman, A. (2005). Sampling for bayesian computation with large datasets. *Available at SSRN 1010107*.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10 :25–37.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An introduction to statistical learning : With applications in R*. Springer.

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., and Bennamoun, M. (2022). Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2) :29–48.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Martin, J. and Elster, C. (2023). Aleatoric uncertainty for errors-in-variables models in deep regression. *Neural Processing Letters*, 55(4) :4799–4818.

# Modèles de Réseaux de neurones avec poids dépendants: Limite, parcimonie et compressibilité

Hoil Lee [1], Fadhel Ayed [2], Paul Jung [3], Juho Lee [1], Hongseok Yang [1], François Caron [4]

[1] *KAIST, South Korea, hoil.lee@kaist.ac.kr, juholee@kaist.ac.kr, hongseok.yang@kaist.ac.kr*
[2] *Huawei Technologies, France, fadhel.ayed@gmail.com*
[3] *Fordham University, USA, pjung3@fordham.edu*
[4] *University of Oxford, UK, caron@stats.ox.ac.uk*

**Résumé.** Ce travail étudie la limite des réseaux neuronaux profonds dont les poids sont dépendants et modélisés via un mélange de distributions gaussiennes. Sous ce modèle, nous montrons que chaque couche du réseau neuronal, losque la largeur tend vers l'infini, peut être caractérisée par deux quantités simples : un paramètre scalaire non-négatif et une mesure de Lévy sur les réels positifs. Si les paramètres scalaires sont strictement positifs et les mesures de Lévy sont triviales alors on retrouve la limite classique du processus gaussien (PG), obtenue avec des poids gaussiens iid. De façon plus intéressante, si la mesure de Lévy d'au moins une couche n'est pas triviale, nous obtenons un mélange de processus gaussiens (MdPG) dans la limite de grande largeur. Le comportement du réseau neuronal dans ce régime est très différent du régime PG. On obtient en effet des sorties corrélées, avec des distributions non gaussiennes, possiblement à queues lourdes. Nous illustrons certains des avantages du régime MdGP sur le régime PG en termes d'apprentissage de représentation et de compressibilité sur des ensembles de données simulées, MNIST et Fashion MNIST.

**Mots-clés.** Réseaux de neurones, processus gaussien, processus stochastique, mesure de Lévy

**Abstract.** This work studies the infinite-width limit of deep feedforward neural networks whose weights are dependent, and modelled via a mixture of Gaussian distributions. Under this model, we show that each layer of the infinite-width neural network can be characterised by two simple quantities: a non-negative scalar parameter and a Lévy measure on the positive reals. If the scalar parameters are strictly positive and the Lévy measures are trivial at all hidden layers, then one recovers the classical Gaussian process (GP) limit, obtained with iid Gaussian weights. More interestingly, if the L'evy measure of at least one layer is non-trivial, we obtain a mixture of Gaussian processes (MoGP) in the large-width limit. The behaviour of the neural network in this regime is very different from the GP regime. One obtains correlated outputs, with non-Gaussian distributions, possibly with heavy tails. We illustrate some of the benefits of the MoGP regime over the GP regime in terms of representation learning and compressibility on simulated, MNIST and Fashion MNIST datasets.

**Keywords.** Neural networks, Gaussian process, stochastic processes, Lévy measure

# 1 Introduction

Deux décennies après le travail fondateur de Radford Neal [1996], ces dernières années ont vu un intérêt renouvelé et croissant pour l'analyse des réseaux neuronaux (profonds) avec des poids aléatoires, dans la limite de largeur infinie. Lorsque les poids sont indépendants et identiquement distribués (iid) selon une loi gaussienne, la fonction aléatoire correspondant à ce réseau neuronal aléatoire converge vers un processus gaussien [Neal, 1996, Lee et al., 2018, Matthews et al., 2018]. La connexion avec les processus gaussiens a approfondi notre compréhension des réseaux neuronaux de grande largeur et a motivé à la fois l'utilisation de méthodes d'inférence de régression bayésienne ou par noyau [Lee et al., 2018] ainsi que le développement de méthodes de noyau pour l'entraînement par descente de gradient dans la limite de largeur infinie [Jacot et al., 2018].

Bien qu'instructive, la connexion avec les processus gaussiens souligne également certaines des limitations des réseaux neuronaux de grande largeur avec des poids gaussiens iid. Comme déjà noté par Neal, " avec des priors gaussiens, les contributions des unités cachées individuelles sont toutes négligeables, et par conséquent, ces unités ne représentent pas des 'caractéristiques cachées' qui capturent des aspects importants des données." De plus, les différentes dimensions de la sortie du réseau neuronal deviennent indépendantes dans la limite de largeur infinie, ce qui est généralement indésirable. Enfin, d'un point de vue bayésien, l'hypothèse d'indépendance gaussienne sur les poids est souvent considérée comme irréaliste : les poids estimés des réseaux neuronaux profonds montrent généralement des dépendances et des queues lourdes, et donc une famille de distributions a priori permettant des queues lourdes est souhaitable. Pour atténuer certaines de ces limitations, des poids aléatoires non gaussiens iid ont été considérés. Cependant, en raison de la même hypothèse iid, certaines des limitations mentionnées persistent, telles que l'indépendance des dimensions de la sortie.

Nous considérons ici une distribution plus structurée sur les poids du réseau neuronal. Nous supposons que les poids émanant d'un noeud donné sont dépendants, et cette dépendance est capturée via un mélange de gaussiennes. Nous nous intéressons à la limite, lorsque la largeur tend vers l'infini, de ce réseau de neurones, et montrons que la limite est un mélange de processus gaussiens. Cette limite est simplement caractérisée, pour chaque couche du réseau, par un paramètre scalaire et une mesure de Lévy. Cette limite a également un certain nombre de propriétés intéressantes par rapport au régime asymptotique standard du processus gaussien: les sorties sont dépendantes, potentiellement avec des queues lourdes, et le réseau de neurone associé est parcimonieux et compressible. Les détails peuvent être trouvés dans l'article [Lee, 2023].

# Bibliographie

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS'18), pages 8571–8580, 2018.

H. Lee, F. Ayed, P. Jung, J. Lee, H. Yang, F. Caron. Deep Neural Networks with Dependent Weights: Gaussian Process Mixture Limit, Heavy Tails, Sparsity and Compressibility. Journal of Machine Learning Research, 2023.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

A. G de G Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

R. M. Neal. Priors for infinite networks. In Bayesian Learning for Neural Networks, pages 29–53. Springer New York, 1996.

# Enquêtes et sondages

# The use of sampling weights in epidemiological research: An application to the KoCo19 study

Ronan Le Gleut [1], Christiane Fuchs [1,2] & the KoCo19 study group

[1] *Core Facility Statistical Consulting, Helmholtz Munich, Germany, ronan.legleut@helmholtz-munich.de*
[2] *Faculty of Business Administration and Economics, Bielefeld University, Germany, christiane.fuchs@uni-bielefeld.de*

**Résumé.** Les études épidémiologiques jouent un rôle central dans la compréhension des maladies, s'appuyant souvent sur des données d'enquêtes pour tirer des conclusions. Malgré les avantages potentiels de la recherche basée sur des enquêtes, l'intégration des poids de sondage, un aspect crucial pour garantir la représentativité des données, reste sous-utilisée dans ce domaine. Les poids de sondage jouent un rôle essentiel dans l'atténuation des biais de non-réponse et l'amélioration de l'exactitude des estimations de paramètres de population calculés à partir de statistiques descriptives. Dans les modèles de régression, les pondérations d'échantillonnage abordent des problèmes tels que l'hétéroscédasticité et le biais, contribuant à la précision et à la fiabilité des estimateurs. Cependant, des défis se posent dans les stratégies de pondération, des hypothèses devant parfois être formulées sur le plan d'échantionnage ou des décisions subjectives devant potentiellement être prises. Il devient alors crucial de trouver un équilibre entre la précision et la simplicité du modèle.

Cet article présente une approche pour calculer des poids dans les études observationnelles, i.e., sans information exhaustive sur le plan de sondage. Elle se rapproche de méthodes existantes utilisant des modèles de superpopulation ou de procédures doublement robuste. La méthode part de poids égaux attribués aux observations et les ajuste par le biais de méthodes de calage utilisant de l'information sociodémographique auxiliaire. Les caractéristiques de l'échantillon sont alors alignées sur les benchmarks de la population connus, corrigeant potentiellement les biais introduits par la non-réponse ou les erreurs de couverture. La méthode peut être particulièrement utile dans les enquêtes complexes où il est difficile d'obtenir un échantillon véritablement aléatoire et représentatif.

L'application des poids (d'échantillonnage) sera démontrée à l'aide de la cohorte représentative sur la COVID-19 à Munich (KoCo19). Cette étude longitudinale, avec un plan de sondage complexe à deux degrés en grappes, étudie la prévalence du virus SARS-CoV-2 et les facteurs de risque associés à une infection.

En conclusion, bien que l'utilisation de poids d'échantillonnage dans les études épidémiologiques soit complexe, elle constitue un élément important afin d'obtenir des statistiques précises et tirer des conclusions valables. En fin de compte, l'application judicieuse de poids (de calage), associée à une communication transparente et à des analyses de sensibilité, est importante afin d'améliorer la fiabilité des conclusions dans la recherche épidémiologique. Ceci est particulièrement important pour une planification et une intervention efficaces en matière de santé publique, où il est vital d'obtenir des taux de prévalence précis et des facteurs de risque fiables.

**Mots-clés.** Études épidémiologiques, Poids de sondage, Méthodes de calage, Représentativité.

**Abstract.** Epidemiological studies play a pivotal role in understanding disease patterns, often relying on survey data to draw meaningful insights. Despite the potential advantages of survey-based research, the incorporation of sampling weights, a crucial aspect of ensuring data representativeness, remains underutilized in this domain. In the realm of descriptive statistics, sampling weights play a vital role in mitigating nonresponse bias and improving the accuracy of population parameter estimates. In regression models, sampling weights address issues like heteroscedasticity and bias, contributing to the precision and reliability of parameter estimates. However, challenges arise in weighting strategies, with assumptions about accurate sampling design information and potential subjective decisions posing hurdles. Striking a balance between precision and model simplicity becomes crucial.

The article introduces an approach for calculating weights in observational studies without exhaustive sampling design information. It is somehow related to superpopulation model or doubly robust procedures. The method starts with equal weights assigned to observations, adjusting them based on auxiliary sociodemographic information through calibration methods. This aligns sample characteristics with known population benchmarks, potentially rectifying biases introduced by nonresponse or coverage errors. The approach might be particularly valuable in complex surveys where obtaining a truly random and representative sample is challenging.

The application of (sampling) weights will be demonstrated using the representative COVID-19 cohort in Munich (KoCo19). This longitudinal population-based study, with a complex two-stage cluster sampling design, investigates SARS-CoV-2 prevalence and risk factors for infection.

In conclusion, the article emphasizes that while the use of sampling weights in epidemiological studies is complex, it is an important component for deriving accurate statistics and making valid conclusions. Ultimately, the judicious application of (calibrated) weights, coupled with transparent reporting and sensitivity analyses, is important for advancing the reliability of epidemiological research findings. This is especially relevant for effective public health planning and intervention, where securing precise prevalence rates and distributions of risk factors is vital.

**Keywords.** Epidemiological studies, Sampling weights, Calibration methods, Representativeness.

# 1 Introduction

Epidemiological studies, pivotal in understanding disease patterns and risk factors, often rely on survey data. However, the use of sampling weights is not necessarily a common practice in such studies, even though they would offer advantages in achieving accurate descriptive statistics and enhancing the validity of regression models. Sampling weights are fundamental for ensuring the representativeness of survey samples, particularly in studies with complex

designs such as stratified or cluster sampling. This is crucial in providing an accurate portrayal of the target population. Using sampling weights may also help to mitigate bias introduced by differential response rates across different subgroups of the population. In regression models, sampling weights play an important role in addressing heteroscedasticity and enhancing the precision and robustness of effect estimates, particularly in epidemiological studies with varying dispersion across subgroups. Despite their advantages, challenges exist in accurately determining the sampling design details, requiring subjective decisions in selecting appropriate weighting strategies and balancing precision with model simplicity to avoid overfitting.

This article aims to present an approach to calculate weights without requiring detailed information about the sampling design, particularly in observational studies. The application of sampling weights will be demonstrated using the representative COVID-19 cohort in Munich (KoCo19).

# 2 Calibration weighting for observational studies

This section presents an approach for calculating weights without requiring detailed information about the sampling design. It is somehow related to the superpopulation model procedure detailed in Valliant, Dever & Kreuter (2018).

Initially, equal weights are assigned to all observations, assuming, e.g., a simple random sampling method. However, this assumption may not hold in many cases, as certain subpopulations may be over- or under-represented. If researchers can incorporate relevant information about the sampling design, it should be utilized to mitigate bias in the analysis, even though this may introduce an element of subjectivity into weight calculations.

To refine survey weights, auxiliary sociodemographic information about the target population is utilized. Calibration methods (Deville & Särndal, 1992) are then applied to align the sample's characteristics with known population totals or benchmarks for specific variables. This calibration technique helps rectify potential biases in survey data arising from nonresponse, coverage errors, or other sampling issues, and only requires to know population totals. Adjusting weights based on known population characteristics aims to enhance the accuracy and reliability of survey estimates, rendering them more representative of the target population. This is particularly valuable in complex surveys where obtaining a truly random and representative sample is challenging.

If it is possible to compute pseudo-inclusion probabilities prior to calibrating the weights, the method can be seen as a doubly robust procedure (Kang & Schafer, 2007). This means that it is approximately unbiased with respect to the quasi-randomization distribution (pseudo inclusion probabilities), to the distribution generated by the superpopulation model (calibration), or to both. However, without any available information on the sampling design, and if the superpopulation model is misspecified, the calibrated estimate may still exhibit bias. Nonetheless, the key questions is whether the bias and the variance have been reduced compared to unweighted estimate.

# 3 Example: The representative COVID-19 cohort Munich (KoCo19)

## 3.1 The KoCo19 cohort

The SARS-CoV-2 virus emerged as a global pandemic in mid-March 2020, just three months after the initial report on December 31, 2019, from the city of Wuhan, Hubei province, China. Seeking a deeper understanding of the actual case numbers, the prospective Munich COVID-19 cohort (KoCo19) was initiated in April 2020, during the initial wave of the pandemic. The cohort comprised 5313 participants aged 13 and above, residing in private households. Over the course of the pandemic, including different waves, variant occurrences, and the commencement of vaccination campaigns, four follow-ups were conducted at critical junctures. The response rates consistently exceeded 70%. In this population-based cohort study, assessment was made on the prevalence of SARS-CoV-2 antibodies (anti-S and anti-N), and participant responses to questionnaires covering sociodemographic information and potential risk factors for infection were collected. Starting from Follow-up 2, information on SARS-CoV-2 vaccination was incorporated.

## 3.2 Sampling design

The participants were selected through a two-stage cluster sampling design.

The first stage involved choosing 100 out of 755 constituencies using a rejective sampling design (Hajek, 1964). Initially, each constituency had an equal probability of being included in the sample ($\sim$13%). The sample of 100 constituencies underwent scrutiny to ensure its representativeness concerning Munich's population in terms of age structure, the percentage of the population with a migration background, households with children, and households with only one member. A sample was deemed representative if the mean fractions in the sample differed from the mean fractions across all 755 Munich constituencies by less than 10 percentage points. Only samples of 100 constituencies meeting these criteria had a non-zero probability of selection. A Monte Carlo simulation with 5000 iterations for random samples of 100 constituencies indicated inclusion probabilities at the constituency level ranging from 12% to 15% (Figure 1A), suggesting that the rejection step did not introduce significant bias.

The second stage involved selecting approximately 30 households for each of the 100 drawn constituencies (Figure 1B), totaling around 3000 households in the sample. These households were obtained through random routes starting in each selected constituency, which could be considered akin to systematic sampling with equal probabilities or simple random sampling. The random routes often crossed constituency borders, allowing a household to be included in the sample via its own constituency or a neighboring one (Figure 1C). To account for these multiple ways of inclusion, we considered first and second-order neighbors (neighbors of a neighbor) for each selected constituency and applied a generalized weight share method (Deville & Lavallée, 2006) for the household weights.

Lastly, all members aged 14 years and older were requested to provide blood samples.

If participants declined to provide blood samples, the sampling weights of the consenting participants within the same household were increased to represent the other members (unit nonresponse treatment within the household).

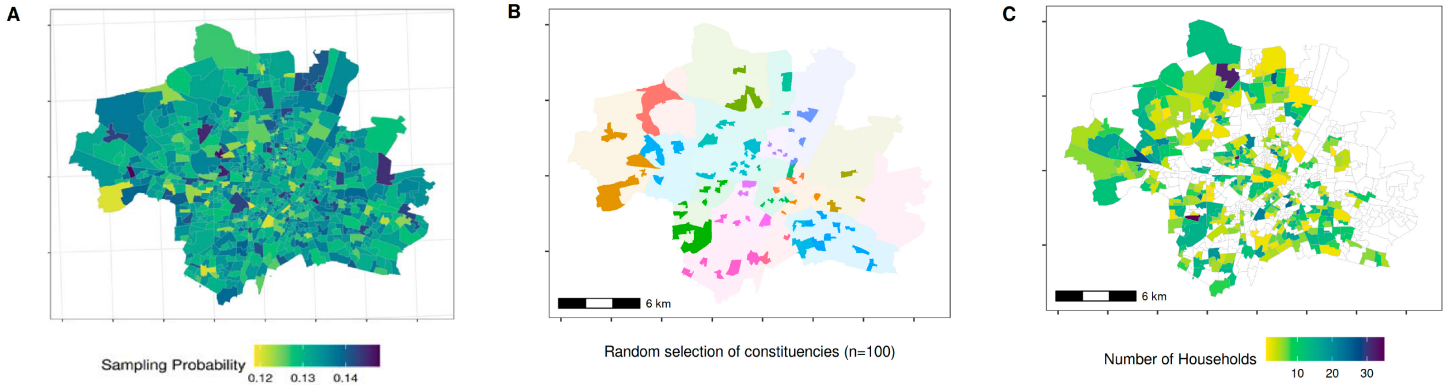



Figure 1: **A** Sampling probabilities for Munich constituencies (rejective sampling) using a Monte Carlo simulation. **B** Munich districts (distinguished by different colors) and the 100 selected start constituencies for the random walks (same color but in a darker shade). **C** All 2994 included households and their respective 368 constituencies.

## 3.3 Post-treatment and variance

After computing sampling weights for all participants, adjustments were made to account for attrition observed at each follow-up, modeling the underlying non-response mechanism (Särndal, Swensson & Wretman, 2003). The resulting weights underwent a final calibration based on the updated structure of the Munich population at each round, considering age, sex, country of birth, presence of children in the household, and single-member household distributions. Under some mild conditions on the sampling design, the variable of interest and the auxiliary variables used in the rejection rule, Fuller (2009) demonstrated that the regression estimator (equivalent to the calibrated estimator) for rejective sampling has similar properties to the regression estimator using the original selection procedure (simple random sampling with inclusion probabilities $\sim$13%), facilitating the calculation of variance and associated confidence intervals.

For the last three follow-ups, information on participants' vaccination status was obtained through questionnaires. Missing values (30% for Follow-up 2, 27% for Follow-up 3, and 8% for Follow-up 4) were imputed via multiple imputation ($m = 100$) using a Bernoulli distribution and crossing vaccination status with information on the immune response (anti-S and anti-N antibodies) for each round. In the last two follow-ups, approximately 93% and 97%, respectively, of participants were assumed vaccinated, in contrast to the city of Munich's reported vaccination rates of approximately 68% and 76% for the population older than 14 years. The calibration of cohort results is therefore of crucial importance.

The variance associated with the calibrated seroprevalence estimates was computed using

linearization and residual techniques (Deville, 1999). Essentially, the variance of the calibrated estimator is asymptotically equivalent to the variance of the total of the residuals of a linear regression using the linearized variable as a response and the auxiliary variables used in the calibration process as covariates. This variance (inference on finite population) accounts for uncertainty due to the different stages of the sampling design (selection of constituencies and households), the non-response mechanism (Juillard & Chauvet, 2018), and the calibration process.

Finally, the variability associated with the multiple imputation procedure was added to the variance of the seroprevalence estimates following the approach detailed in Honaker, King & Blackwell (2011). In short, the final variance estimate $V$ is a combination of the average of the variance estimates $V_j,\ j = 1, \ldots, m$ (described above) over $m$ replications and the variance of the $m$ seroprevalence estimates $\theta_j,\ j = 1, \ldots, m$:

$$V = \frac{1}{m}\sum_{j=1}^{m} V_j + S^2\left(1 + \frac{1}{m}\right), \text{ with } S^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(\theta_j - \overline{\theta}\right)^2$$

The final seroprevalence estimates were obtained using the means of the $m$ estimates, and 95% confidence intervals were computed assuming a normal distribution. Chauvet & Vallée (2020) have given general conditions for the asymptotic normality of the Horvitz-Thompson estimator under two-stage sampling designs.

In addition to accounting for the sampling design, the probabilities of the laboratory tests to yield false negatives or false positive results were considered. Following Sempos & Tian (2021), adjusted seroprevalence was calculated as $(\theta + sp - 1) \,/\, (sen + sp - 1)$, where $sp$ represents the estimated specificity and $sen$ represents the estimated sensitivity. While this adjustment is an exact formula for true seroprevalence, sensitivity and specificity, it becomes only approximate if the estimates are calculated and plugged in independently.

## 3.4 Unweighted estimates

As a sensitivity analysis, unweighted seroprevalence estimates were also computed along with their uncertainty. The variance was determined by a nonparametric cluster bootstrap procedure accounting for household clustering (Cameron, Gelbach & Miller, 2008). Seroprevalence estimates were calculated in each of the 5000 bootstrap samples (sampling households with replacement), and the variance of these estimates provided the uncertainty of the unweighted estimates. These seroprevalence estimates were also adjusted for the sensitivity and specificity of the laboratory tests.

## 3.5 Results

The weighted (calibrated) cumulative seroprevalence, adjusted for sensitivity and specificity, in private households for the Munich population aged 14 years and older increased from 1.6% (1.1-2.1%) in March 2020 to 12.4% (10.7-14.1%) in August 2021 and 14.5% (12.7-

16.2%) in November 2021 (Figure 2A). Without adjusting for vaccination status in Follow-ups 3 and 4, the seroprevalence would have been significantly lower: 8.5% (7.2-9.8%) for August 2021 and 10.5% (9.1-11.9%) for November 2021. Indeed, the proportion of vaccinated individuals is higher in the cohort compared to the general Munich population. Therefore, calibrating for vaccination status increases the weight of participants who are not vaccinated. As the seroprevalence is higher in the non-vaccinated population (Figure 2B), the overall seroprevalence, encompassing both vaccinated and non-vaccinated individuals, also increases with the calibration.

As a sensitivity analysis, unweighted estimates are presented in green (Figure 2A). Weighted estimates without calibration for vaccination status (in orange) and unweighted estimates are very close. Despite some weight sharing and nonresponse, the sample was already representative of the Munich population, or its characteristics do not influence the infection status (except for vaccination status).

The official number of positive cases is indicated in pink (Figure 2A) for the general population of Munich. In contrast to the KoCo19 cohort, this figure incorporates institutions such as nursing homes and encompasses potential cases of reinfection. Since the KoCo19 cohort is limited to private households and that the estimated seroprevalence does not account for multiple infections (neglected before the Omicron variant), comparing this estimate with the official number over time allows us to estimate a lower bound for the underreporting factor. The estimated underreporting factor changed over the rounds from 3.4 (2.4–4.4) at Baseline to 2.2 (2.0-2.5) at Follow-up 4.

Additional findings regarding sero-incidence, breakthrough infections, and infections among naïve subjects are available in Le Gleut et al. (2023).

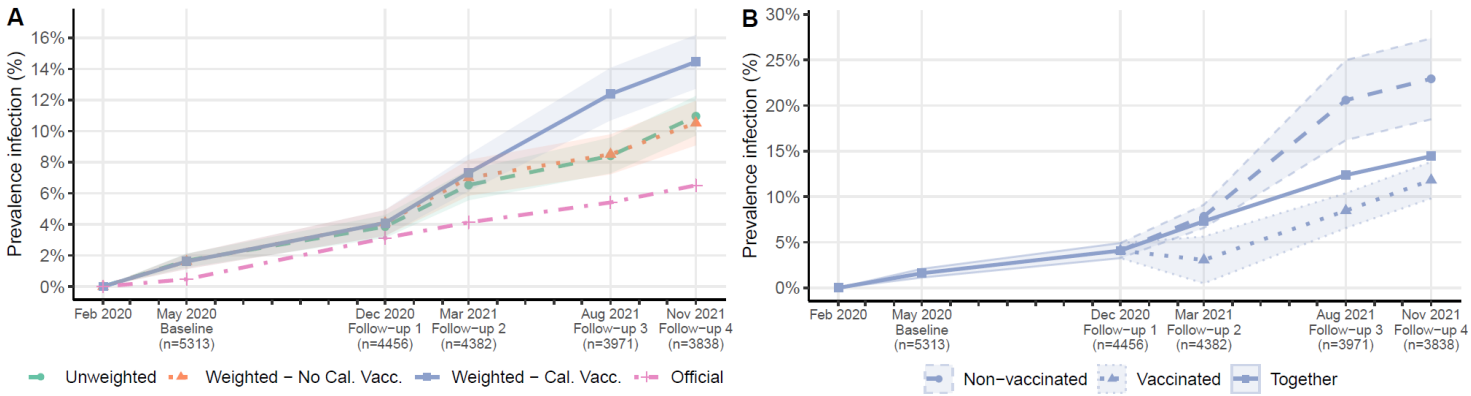



Figure 2: **A** Weighted and unweighted cumulative seroprevalence in private households and official numbers of cases reported by the authorities for the Munich population older than 13 years. **B** Seroprevalence estimates calibrated on the number of vaccinated people split according to the vaccination status of the same round.

Throughout the entire follow-up period, factors such as being born outside Germany, working in a high-risk job, and the area of residence per inhabitant were identified as infection risk factors, while other sociodemographic and health-related variables were not significant

(Le Gleut et al., 2023). This analysis utilized an extended Cox regression model (Anderson & Gill, 1982; Therneau & Grambsch, 2000) to account for intra-household clustering in the data and to obtain robust standard error estimates. During the examination of risk factors, the consideration of weights was omitted. However, in line with the concepts introduced by Solon, Haider, and Wooldridge (2015), one could have conducted appropriate diagnostics before determining the justification for using weights. A sensitivity analysis would have been pertinent as well, where the comparison between weighted and unweighted estimates (Wooldridge, 2001) could serve as a diagnostic tool for model misspecification or endogenous sampling.

Significantly more outcomes emerged from this cohort, encompassing a head-to-head evaluation of various seroassays (Olbrich et al., 2021), an integrative modeling approach to reported case numbers and seroprevalence utilizing a compartment model (Contento et al., 2023), and an evaluation of T cell reactivity following SARS-CoV-2 infection (Brand et al., 2021).

# 4 Discussion

In conclusion, the use of sampling weights in epidemiological studies is a complex endeavor with both advantages and challenges. While essential for achieving accurate descriptive statistics and enhancing the validity of statistical inferences, researchers must navigate the potential uncertainties associated with sampling design and subjective choices in weighting strategies. The judicious application of sampling weights, coupled with transparency in reporting, sensitivity analyses, and an understanding of study design, is essential to harness the full benefits of this methodology.

# References

Brand, I., Gilberg, L., Bruger, J., Garí, M., Wieser, A., Eser, T. M., ... & Geldmacher, C. (2021). Broad T cell targeting of structural proteins after SARS-CoV-2 infection: High throughput assessment of T cell reactivity using an automated interferon gamma release assay. *Frontiers in Immunology*, 12, 688436.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics*, 90(3), 414-427.

Chauvet, G., & Vallée, A.A. (2020). Consistency of estimators and variance estimators in two-stage sampling. *J. Royal Stat. Soc.*, 82, 797–815.

Contento, L., Castelletti, N., Raimúndez, E., Le Gleut, R., Schälte, Y., Stapor, P., ... & KoCo19 study group. (2023). Integrative modelling of reported case numbers and seroprevalence reveals time-dependent test efficiency and infectious contacts. *Epidemics*, 43, 100681.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey methodology*, 25(2), 193-204.

Deville, J., & Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32(2), 165.

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418), 376-382.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4), 1491-1523.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software*, 45, 1-47.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, 22(4), 523 - 539.

Le Gleut, R., Plank, M., Pütz, P., Radon, K., Bakuli, A., Rubio-Acero, R., ... & Castelletti, N. (2023). The representative COVID-19 cohort Munich (KoCo19): from the beginning of the pandemic to the Delta virus variant. *BMC Infectious Diseases*, 23(1), 466.

Olbrich, L., Castelletti, N., Schälte, Y., Garí, M., Pütz, P., Bakuli, A., ... & KoCo19-Study Group. (2021). Head-to-head evaluation of seven different seroassays including direct viral neutralisation in a representative cohort for SARS-CoV-2. *Journal of General Virology*, 102(10), 001653.

Särndal, C. E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. *Springer Science & Business Media*.

Sempos, C. T., & Tian, L. (2021). Adjusting coronavirus prevalence estimates for laboratory test kit error. *American journal of epidemiology*, 190(1), 109-115.

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What Are We Weighting For? *Journal of Human Resources*, 50(2), 301-316.

Valliant, R., Dever, J., & Kreuter, F. (2018). Practical Tools for Designing and Weighting Survey Samples. *Springer*.

Wooldridge, J. M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric theory*, 17(2), 451-470.

# Asymptotic properties of estimators for continuous sampling designs with application to environmental surveys

Guillaume Chauvet [1] & Minna Pulkkinen [2]

[1] *Univ. Rennes, ENSAI, CNRS, IRMAR–UMR 6625, F-35000, Rennes, France, chauvet@ensai.fr*
[2] *IGN, ENSG, Laboratoire d'Inventaire Forestier (LIF), Nancy, France, minna.pulkkinen@ign.fr*

**Résumé.** Les inventaires forestiers nationaux sont basés sur des plans de sondage probabilistes. Il est courant de sélectionner aléatoirement un échantillon de points dans un continuum (le territoire étudié) puis de définir des supports de forme fixe (par exemple, des placettes ou des polygones) à partir de ces points pour réaliser l'enquête sur la population d'arbres ; voir par exemple Vidal et al. (2016) pour un aperçu des plans d'échantillonnage utilisés. Bien que le plan d'échantillonnage puisse être formalisé de plusieurs manières (voir par exemple Eriksson, 1995), l'approche de la population infinie (Stevens et Urqhart, 2000; Barabesi, 2003; Mandallaz, 2007; Grégoire et Valentine, 2007) est sans doute le dispositif le plus simple. L'inférence peut être effectuée directement à partir de la population échantillonnée, en utilisant la théorie de l'estimation de Horvitz-Thompson en population continue (Cordy, 1993), à la fois en termes d'estimation ponctuelle et d'estimation de variance (Chauvet et al., 2023). Afin de produire des estimateurs fiables, certaines propriétés importantes sont nécessaires à ces plans d'échantillonnage. L'estimateur de Horvitz-Thompson doit être consistant et asymptotiquement normalement distribué, et un estimateur de variance consistant doit également être disponible pour une estimation par intervalle. Ces propriétés ont été peu étudiées dans la littérature, à l'exception de Barebesi et Franceschi (2011) et Barabesi et al. (2012). Dans ce travail, nous obtenons ces propriétés de façon générale pour des plans de sondage en population continue, sous des conditions faibles.

**Mots-clés.** Consistance, Enquête environnementale, Inventaire forestier, Normalité asymptotique.

**Abstract.** National forest inventories are based on probabilistic sampling designs. It is common practice to randomly select a sample of points in a continuum (the territory under study) and then to define fixed-shape supports (e.g., plots or polygons) from these points to perform the survey on the field on the population of trees; see for example Vidal et al. (2016) for a worldwide overview of sampling designs used in forest inventories. Although the sampling design may be formalized in several manners (e.g., Eriksson, 1995), the infinite population approach (Stevens and Urqhart, 2000; Barabesi, 2003; Mandallaz, 2007; Gregoire and Valentine, 2007) is arguably the simplest device for inference. Inference may be performed directly from the sampled population, which is straightforward by using the theory of continuous Horvitz-Thompson (HT) estimation (Cordy, 1993), both in terms of point estimation and variance estimation (Chauvet et al., 2023). In order to produce reliable estimators, some important properties are needed for continuous sampling designs. The

HT-estimator should be consistent and asymptotically normally distributed, with consistent variance estimators for interval estimation. These properties have not been much considered in the literature, with the exception of Barebesi and Franceschi (2011) and Barabesi et al. (2012). In this work, we derive these properties for general continuous sampling designs, under mild conditions.

**Keywords.** Consistency, Environmental survey, Forest inventory, Asymptotic normality.

# 1 Continuous Horvitz-Thompson estimation

## 1.1 Notations

We are interested in a continuous universe $\mathcal{U}$ of surface area $A$, which is included in $\mathbb{R}^q$ for some $q \geq 1$. We consider some $Q$-vector of attributes $\boldsymbol{\rho} : \mathcal{U} \to \mathbb{R}^Q$ which is Lebesgue-integrable, and we wish to estimate the (integral) total

$$\boldsymbol{\tau}_{\boldsymbol{\rho}} \quad = \quad \int_{\mathcal{U}} \boldsymbol{\rho}(x) \, \mathrm{d}x. \tag{1}$$

In case of forest inventories, $\mathcal{U}$ is an intermediary universe used to attain a population $U$ of trees for which we want to estimate the total $\boldsymbol{Y} = \sum_{k \in U} \boldsymbol{y}_k$ of some $Q$-vector of interest $\boldsymbol{y}_k$. Making use of a non-negative link function $L(\cdot, \cdot) : \mathcal{U} \times U \to \mathbb{R}^+$ which depends on the form of the fixed-shape support, the attribute of interest $\boldsymbol{y}_k$ may be converted into the synthetic variable

$$\boldsymbol{\rho}(x) = \sum_{k \in U} \frac{L(x,k)\boldsymbol{y}_k}{M_{+k}} \text{ for } x \in \mathcal{U}, \tag{2}$$

with $M_{+k} = \int_{\mathcal{U}} L(x,k) \, \mathrm{d}x$ the measure of the sub-territory in $\mathcal{U}$ leading to the selection of unit $k$, see Stevens and Urquart (2000) and Chauvet et al. (2023). Provided that $M_{+k} > 0$ for any $k \in U$, the local variable $\boldsymbol{\rho}(x)$ is such $\int_{\mathcal{U}} \boldsymbol{y}(x) dx = \boldsymbol{Y}$, hence estimating the total $\boldsymbol{Y}$ is equivalent to estimate the integral in (1).

A random sample $S = \{x_1, \ldots, x_n\}$ of $n$ locations is selected in $\mathcal{U}$ according to the joint probability density function (PDF) $f(x_1, \ldots, x_n)$. Following Cordy (1993), we define the inclusion density function for $x \in \mathcal{U}$ as

$$\pi(x) \quad = \quad \sum_{i=1}^{n} f_i(x), \tag{3}$$

where $f_i(\cdot)$ if the marginal PDF for the $i^{\text{th}}$ draw, and we define the joint inclusion density function for $x, y \in \mathcal{U}^2$ as

$$\pi(x,y) \quad = \quad \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} f_{ij}(x,y), \tag{4}$$

where $f_{ij}(\cdot,\cdot)$ if the joint PDF for the draws $i$ and $j$. The third order and fourth order inclusion density functions

$$\begin{aligned} \pi(x,y,z) &= \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{i'=1\\ i'\neq i,j}}^{n} f_{iji'}(x,y,z), \qquad (5)\\ \text{and } \pi(x,y,z,t) &= \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{i'=1\\ i'\neq i,j}}^{n}\sum_{\substack{j'=1\\ j'\neq i,j,i'}}^{n} f_{iji'j'}(x,y,z,t) \end{aligned}$$

are defined similarly, with $f_{iji'}(\cdot,\cdot,\cdot)$ the joint PDF for the draws $i$, $j$, $j'$ and $f_{iji'j'}(\cdot,\cdot,\cdot,\cdot)$ the joint PDF for the draws $i$, $j$, $j'$ and $i'$.

## 1.2 Horvitz-Thompson estimation

The Horvitz-Thompson (HT) estimator

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi} = \sum_{x\in S}\frac{\boldsymbol{\rho}(x)}{\pi(x)} = \sum_{i=1}^{n}\frac{\boldsymbol{\rho}(x_i)}{\pi(x_i)} \qquad (6)$$

is design-unbiased for $\boldsymbol{\tau}_{\boldsymbol{\rho}}$, provided that $\pi(x) > 0$ almost everywhere on $\mathcal{U}$ (Cordy, 1993). The covariance matrix of $\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}$ is

$$\boldsymbol{V}_p(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) = \int_{\mathcal{U}}\frac{\boldsymbol{\rho}(x)\boldsymbol{\rho}(x)^\top}{\pi(x)}\,\mathrm{d}x + \int_{\mathcal{U}}\int_{\mathcal{U}}\{\pi(x,y)-\pi(x)\pi(y)\}\frac{\boldsymbol{\rho}(x)\boldsymbol{\rho}(y)^\top}{\pi(x)\pi(y)}\,\mathrm{d}x\mathrm{d}y, \qquad (7)$$

and may be estimated by the Horvitz-Thompson variance estimator

$$\hat{\boldsymbol{V}}_{HT}(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) = \sum_{i=1}^{n}\frac{\boldsymbol{\rho}(x_i)\boldsymbol{\rho}(x_i)^\top}{\pi(x_i)^2} + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{\pi(x_i,x_j)-\pi(x_i)\pi(x_j)}{\pi(x_i,x_j)}\frac{\boldsymbol{\rho}(x_i)}{\pi(x_i)}\frac{\boldsymbol{\rho}(x_j)^\top}{\pi(x_j)}, \qquad (8)$$

and the Sen-Yates-Grundy variance estimator

$$\hat{\boldsymbol{V}}_{YG}(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{\pi(x_i)\pi(x_j)-\pi(x_i,x_j)}{\pi(x_i,x_j)}\left\{\frac{\boldsymbol{\rho}(x_i)}{\pi(x_i)}-\frac{\boldsymbol{\rho}(x_j)}{\pi(x_j)}\right\}\left\{\frac{\boldsymbol{\rho}(x_i)}{\pi(x_i)}-\frac{\boldsymbol{\rho}(x_j)}{\pi(x_j)}\right\}^\top. \qquad (9)$$

Cordy (1993) considered only the case $Q = 1$, but his variance formulas are easily generalized to the multivariate case. For the fixed-size sampling designs that we consider in this paper, the two variance estimators are design-unbiased for $\boldsymbol{V}_p(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi})$, provided that $\pi(x,y) > 0$ almost everywhere on $\mathcal{U}^2$ (Cordy, 1993, Section 2).

## 1.3 Plug-in estimation

We also consider the case of smooth functions of totals, which is very important in practice. Suppose that we wish to estimate a parameter $\theta = g(\boldsymbol{\tau}_{\boldsymbol{\rho}})$, with $g : \mathbb{R}^Q \to \mathbb{R}$ some known function. Let $\|\cdot\|$ denote the Euclidean norm. Let us denote

$$\boldsymbol{\mu}_{\boldsymbol{\rho}} = \frac{\boldsymbol{\tau}_{\boldsymbol{\rho}}}{A} \quad \text{and} \quad \hat{\mu}_{\boldsymbol{\rho}\pi} = \frac{\hat{\tau}_{\boldsymbol{\rho}\pi}}{A}. \tag{10}$$

We suppose that there exists some neighborhood $B_\eta(\boldsymbol{\mu}_{\boldsymbol{\rho}}) = \{\mathbf{a} \in \mathbb{R}^q;\ \|\mathbf{a} - \boldsymbol{\mu}_{\boldsymbol{\rho}}\| \leq \eta\}$ of $\boldsymbol{\mu}_{\boldsymbol{\rho}}$, for some $\eta > 0$, such that $g(\cdot)$ is differentiable on $B_\eta(\boldsymbol{\mu}_{\boldsymbol{\rho}})$, with $\boldsymbol{g}'(\boldsymbol{\mu}_{\boldsymbol{\rho}}) \neq 0$. The plug-in estimator of $\theta$, truncated to avoid impossible values, is

$$\hat{\theta}_\pi \quad = \quad \begin{cases} g(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) & \text{if } \hat{\boldsymbol{\mu}}_{\boldsymbol{\rho}\pi} \in B_\eta(\boldsymbol{\mu}_{\boldsymbol{\rho}}), \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The linearization variance approximation for $\hat{\theta}_\pi$ is $V_p(\hat{\tau}_{l\pi})$, with

$$l(x) \quad = \quad \{\boldsymbol{g}'(\boldsymbol{\tau}_{\boldsymbol{\rho}})\}^\top \boldsymbol{\rho}(x), \tag{12}$$

the linearized variable of $\theta$, and where $\hat{\tau}_{l\pi}$ is obtained from (6) by replacing $\boldsymbol{\rho}(x)$ with $l(x)$. The truncated Sen-Yates-Grundy linearized variance estimator is obtained by replacing in equation (9) the variable $\boldsymbol{\rho}(x)$ with the estimated linearized variable

$$\hat{l}(x) \quad = \quad \{\boldsymbol{g}'(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi})\}^\top \boldsymbol{\rho}(x), \tag{13}$$

which leads to

$$\hat{V}_{YG}(\hat{\theta}_\pi) = \begin{cases} \frac{1}{2}\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \frac{\pi(x_i)\pi(x_j)-\pi(x_i,x_j)}{\pi(x_i,x_j)} \left\{\frac{\hat{l}(x_i)}{\pi(x_i)} - \frac{\hat{l}(x_j)}{\pi(x_j)}\right\}^2 & \text{if } \hat{\boldsymbol{\mu}}_{\boldsymbol{\rho}\pi} \in B_\eta(\boldsymbol{\mu}_{\boldsymbol{\rho}}), \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The Horvitz-Thompson linearized variance estimator

$$\hat{V}_{HT}(\hat{\theta}_\pi) = \begin{cases} \sum_{i=1}^n \left\{\frac{\hat{l}(x_i)}{\pi(x_i)}\right\}^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \frac{\pi(x_i,x_j)-\pi(x_i)\pi(x_j)}{\pi(x_i,x_j)} \frac{\hat{l}(x_i)}{\pi(x_i)} \frac{\hat{l}(x_j)}{\pi(x_j)} & \text{if } \hat{\boldsymbol{\mu}}_{\boldsymbol{\rho}\pi} \in B_\eta(\boldsymbol{\mu}_{\boldsymbol{\rho}}), \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

is obtained similarly.

## 1.4 General assumptions

The asymptotic framework for finite population sampling usually postulates that the population of interest belongs to a nested sequence of populations with increasing sizes (Isaki and Fuller, 1982), since if the sample size grows to infinity, the population size needs to grow accordingly. The asymptotic framework with continuous sampling designs is simpler, since there is no contradiction in having an increasing sample size inside an infinite universe (Mandallaz, 2007, p. 61). Therefore, we use the framework introduced in Mandallaz (1991),

and simply suppose that the population $\mathcal{U}$ remains fixed, and that $n \to \infty$. This is also known as the Monte Carlo integration approach (Gregoire and Valentine, 2007, chapter 10).

We also consider the following assumptions:

H1: Some constant $M_1$ exists such that

$$\frac{1}{A}\int_{\mathcal{U}} \|\boldsymbol{\rho}(x)\|^4 \,\mathrm{d}x \;\;\leq\;\; M_1.$$

H2: Some constants $c_1, C_1 > 0$ exist such that for any $x, y \in \mathcal{U}^2$:

$$c_1\frac{n^2}{A^2} \leq \pi(x,y) \leq C_1\frac{n^2}{A^2}.$$

H3: Some constant $C_2$ exists such that

$$\sup_{x,y\in\mathcal{U}^2} |\Delta_2(x,y)| \leq C_2\frac{n}{A^2}, \tag{16}$$

with $\Delta_2(x,y) = \pi(x,y) - \pi(x)\pi(y)$. Some constant $C_3$ exists such that

$$\sup_{x,y,z,t\in\mathcal{U}^4} |H_4(x,y,z,t)| \leq C_3\frac{n^2}{A^4}, \tag{17}$$

where

$$\begin{aligned} H_4(x,y,z,t) &= \Delta_4(x,y,z,t) \\ &- \{\pi(x)\Delta_3(y,z,t) + \pi(y)\Delta_3(x,z,t) + \pi(z)\Delta_3(x,y,t) + \pi(t)\Delta_3(x,y,z)\} \\ &+ \{\,\pi(z)\pi(t)\Delta_2(x,y) + \pi(y)\pi(t)\Delta_2(x,z) + \pi(y)\pi(z)\Delta_2(x,t) \\ &\quad +\pi(x)\pi(t)\Delta_2(y,z) + \pi(x)\pi(z)\Delta_2(y,t) + \pi(x)\pi(y)\Delta_2(z,t)\} \end{aligned}$$

and with

$$\begin{aligned} \Delta_3(x,y,z) &= \pi(x,y,z) - \pi(x)\pi(y)\pi(z), \\ \Delta_4(x,y,z,t) &= \pi(x,y,z,t) - \pi(x)\pi(y)\pi(z)\pi(t). \end{aligned}$$

H4: $g(\cdot)$ is homogeneous of degree $\beta \geq 0$, i.e. $g(r\mathbf{a}) = r^\beta g(\mathbf{a})$ for any real $r > 0$ and any $\mathbf{a} \in \mathbb{R}^Q$.

H5: The differential $\boldsymbol{g}'$ is locally Lipschitz on $B_\eta(\mu_{\boldsymbol{\nu}})$, i.e. there exists some constant $K$ such that for any $Q$-vectors $\mathbf{a}, \mathbf{b} \in B_\eta(\mu_{\boldsymbol{\nu}}) \times B_\eta(\mu_{\boldsymbol{\nu}})$, we have $\|\boldsymbol{g}'(\mathbf{b}) - \boldsymbol{g}'(\mathbf{a})\| \leq K\|\mathbf{b} - \mathbf{a}\|$.

It is assumed in (H1) that the vector of interest $\rho(x)$ has a finite moment of order 4, which holds in particular if all the components of $\rho(x)$ are bounded. This is a fairly weak assumption. Assumption (H2) is related to the joint inclusion density function, which is assumed to be bounded both above and below by the joint inclusion density function obtained under uniform

random sampling. Assumption (H3) is related to the inclusion density functions of order 2 to 4. This is the equivalent of classical assumptions for finite population sampling, see for example Breidt and Opsomer (2000), assumptions (A6) and (A7). We show in Section 2 that assumptions (H2) and (H3) are satisfied in case of stratified sampling in $\mathcal{U}$ with a finite number of strata. The assumption (H4) was introduced in Deville (1999), and is satisfied with $\beta = 0$ for many parameters of interest like a ratio or a correlation coefficient, for example. The assumption (H5) is a mild regularity condition for $g(\cdot)$ in the neighborhood of $\boldsymbol{\mu_\rho}$. The condition of Lipschitz continuity is satisfied if $g(\cdot)$ is twice differentiable with a bounded Hessian matrix on $B_\eta(\boldsymbol{\mu_\rho})$.

## 1.5 Consistency of estimators

We first obtain in Proposition 1 the $\sqrt{n}$-consistency of the HT-estimators, and of the associated estimators of the covariance matrix.

**Proposition 1.** *Suppose that assumptions (H1)-(H3) hold. Then:*

$$\left\| V\left\{A^{-1}\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi} - \boldsymbol{\tau}_{\boldsymbol{\rho}}\right)\right\}\right\| = O(n^{-1}), \tag{18}$$

$$E\left[\left\|A^{-2}n\left\{\hat{\boldsymbol{V}}_{HT}(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) - \boldsymbol{V}_p(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi})\right\}\right\|^2\right] = O(n^{-1}), \tag{19}$$

$$E\left[\left\|A^{-2}n\left\{\hat{\boldsymbol{V}}_{YG}(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) - \boldsymbol{V}_p(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi})\right\}\right\|^2\right] = O(n^{-1}), \tag{20}$$

*where for a square matrix $\boldsymbol{B}$, $\|\boldsymbol{B}\|$ is the matrix norm induced by the Euclidean norm, a.k.a. the spectral norm.*

Proposition 1 ensures that the HT-estimator is mean-square consistent, and that both the HT estimator and the YG estimator are also mean-square consistent for the covariance matrix, and may therefore be used for interval estimation. In the context of a forest inventory in a territory $\mathcal{U}$, Proposition 1 is of interest for both the estimation of surface attributes and of tree attributes. For a surface attribute, we may for example be interested in the domain $\mathcal{U}_d$ where forest is located. The corresponding area $A_d$ may be written as the integral in (1), with $Q = 1$ and $\rho(x) = 1(x \in \mathcal{U}_d)$ the domain membership indicator. For a tree attribute, we may for example be interested in the total volume of wood $Y = \sum_{k\in U} y_k$, with $U$ the population of trees located in $\mathcal{U}$, and $y_k$ the volume of tree $k$. Making use of the synthetic variable in (2), $Y$ may also we written as the integral in (1). Proposition 1 is therefore applicable for these two types of parameters.

**Proposition 2.** *Suppose that assumptions (H1)-(H5) hold. Then:*

$$E\left(A^{-2\beta}\left[(\hat{\theta}_\pi - \theta) - \{g'(\tau_{\boldsymbol{\rho}})\}^\top\{\hat{\tau}_{\boldsymbol{\rho}\pi} - \tau_{\boldsymbol{\rho}}\}\right]\right)^2 = O(n^{-2}), \tag{21}$$

$$E\left[A^{-2\beta}n\left|\hat{V}_{HT}(\hat{\theta}_\pi) - V_p(\hat{\tau}_{l\pi})\right|\right] = O(n^{-1/2}), \tag{22}$$

$$E\left[A^{-2\beta}n\left|\hat{V}_{YG}(\hat{\theta}_\pi) - V_p(\hat{\tau}_{l\pi})\right|\right] = O(n^{-1/2}). \tag{23}$$

Equation (21) in Proposition 2 implies that the bias of the truncated plug-in estimator is asymptotically negligible. Equation (21) also implies that the linearization variance approximation and the true variance of the truncated plug-in estimator are asymptotically the same.

The consistency in the $L^1$ norm of the HT and SYG variance estimators is established in equations (22) and (23), which is weaker than the mean-square consistency obtained in Proposition 1 for the estimators of the covariance matrix of $\hat{\tau}_{\boldsymbol{\rho}\pi}$. The mean square consistency can easily be established by using the same proof, and by strengthening (H1) to have a bounded moment of order 8. Anyway, the convergence in the $L^1$ norm of the variance estimators is sufficient to obtain the asymptotic normality of the studentized plug-in estimator in case of stratified sampling, see Section 2.

# 2 Stratified sampling

## 2.1 Notations

The population $\mathcal{U}$ is partitioned into $H$ strata $\mathcal{U}^1, \ldots, \mathcal{U}^H$ with respective areas $A^1, \ldots, A^H$. Inside the stratum $\mathcal{U}^h$, a sample $S^h$ is obtained by $n^h$ independent selections according to the marginal PDF $g^h(\cdot)$, and

$$S \;=\; \bigcup_{h=1}^{H} S^h. \tag{24}$$

The integral total $\boldsymbol{\tau}_{\boldsymbol{\rho}}$ in (1) may we rewritten as

$$\boldsymbol{\tau}_{\boldsymbol{\rho}} = \sum_{h=1}^{H} \boldsymbol{\tau}_{\boldsymbol{\rho}}^h \quad \text{with} \quad \boldsymbol{\tau}_{\boldsymbol{\rho}}^h = \int_{\mathcal{U}^h} \boldsymbol{\rho}(x)\,\mathrm{d}x.$$

From equation (3), we have

$$\pi(x) \;=\; n^h g^h(x) \;\text{ for any }\; x \in \mathcal{U}^h, \tag{25}$$

and the HT-estimator may be rewritten as

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi} \;=\; \sum_{h=1}^{H} \hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}^h \text{ with } \hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}^h = \frac{1}{n^h} \sum_{x \in S^h} \frac{\boldsymbol{\rho}(x)}{g^h(x)}. \tag{26}$$

From equation (4), we have

$$\pi(x,y) \;=\; \begin{cases} n^h(n^h-1)g^h(x)g^h(y) & \text{if } x,y \in \mathcal{U}^h, \\ n^h n^{h'} g^h(x) g^{h'}(y) & \text{if } x \in \mathcal{U}^h \text{ and } y \in \mathcal{U}^{h'} \text{ with } h \neq h'. \end{cases} \tag{27}$$

In particular, note that we need $n^h \geq 2$ for any $h = 1, \ldots, H$ for $\pi(x, y)$ to be positive on $\mathcal{U}^2$. By using equation (5), the covariance matrix of $\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}$ may be rewritten as

$$\begin{aligned} \boldsymbol{V}_p(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) &= \sum_{h=1}^{H} \frac{\boldsymbol{\Sigma}^2_{\boldsymbol{\rho}h}}{n^h}, \\ \text{with } \boldsymbol{\Sigma}^2_{\boldsymbol{\rho}h} &= \int_{\mathcal{U}^h} g^h(x) \left\{ \frac{\boldsymbol{\rho}(x)}{g^h(x)} - \boldsymbol{\tau}^h_{\boldsymbol{\rho}} \right\} \left\{ \frac{\boldsymbol{\rho}(x)}{g^h(x)} - \boldsymbol{\tau}^h_{\boldsymbol{\rho}} \right\}^\top \mathrm{d}x. \end{aligned} \tag{28}$$

Suppose that $n^h \geq 2$ for any $h = 1, \ldots, H$. The Sen-Yates-Grundy variance estimator may be rewritten as

$$\begin{aligned} \hat{\boldsymbol{V}}_{YG}(\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi}) &= \sum_{h=1}^{H} \frac{\hat{\boldsymbol{\Sigma}}^2_{\boldsymbol{\rho}h}}{n^h}, \\ \text{with } \hat{\boldsymbol{\Sigma}}^2_{\boldsymbol{\rho}h} &= \frac{1}{n^h - 1} \sum_{x \in S^h} \left\{ \frac{\boldsymbol{\rho}(x)}{g^h(x)} - \hat{\boldsymbol{\tau}}^h_{\boldsymbol{\rho}\pi} \right\} \left\{ \frac{\boldsymbol{\rho}(x)}{g^h(x)} - \hat{\boldsymbol{\tau}}^h_{\boldsymbol{\rho}\pi} \right\}^\top, \end{aligned} \tag{29}$$

and the Horvitz-Thompson variance estimator is identical.

## 2.2 Assumptions

We consider the following assumptions for stratified sampling:

H1b: There exists some constant $M_2 > 0$ such that:

$$\sup_{h=1,\ldots,H} \frac{1}{A^h} \int_{\mathcal{U}_h} \|\boldsymbol{\rho}(x)\|^4 \, \mathrm{d}x \leq M_2.$$

H6: There exists some constants $c_5, C_5 > 0$ such that for any $h = 1, \ldots, H$ and any $x \in \mathcal{U}_h$:

$$c_5 \leq A^h g^h(x) \leq C_5.$$

H7: There exists some constants $c_6, C_6 > 0$ and $C_7, C_7 > 0$ such that for any $h = 1, \ldots, H$:

$$\begin{aligned} c_6 \frac{n}{H} &\leq n_h \leq C_6 \frac{n}{H}, \\ c_7 \frac{A}{H} &\leq A^h \leq C_7 \frac{A}{H}. \end{aligned}$$

It is assumed in (H1b) that the vector of interest $\boldsymbol{\rho}(x)$ has a finite moment of order 4 inside each stratum. It is assumed in (H6) that the probability density function is of the same order for all the points inside a given stratum. In particular, this assumption holds true in case of uniform sampling inside strata. It is assumed in (H7) that the sample size share is of same order for each stratum, and similarly that the surface area share is of same order for each stratum.

**Theorem 1.** *Suppose that $S$ is selected by stratified sampling, with a finite number of strata $H$ and $n^h \geq 2$ inside each stratum $\mathcal{U}^h$. Suppose that assumptions (H1b), (H6) and (H7) hold. Then assumptions (H1)-(H3) hold, and the conclusions of Proposition 1 hold true. Also, for any $h = 1, \ldots, H$:*

$$\sqrt{n^h}\left\{\hat{\boldsymbol{\tau}}^h_{\boldsymbol{\rho}\pi} - \boldsymbol{\tau}^h_{\boldsymbol{\rho}}\right\} \quad \longrightarrow_{\mathcal{L}} \quad \mathcal{N}\left(\mathbf{0}_Q, \boldsymbol{\Sigma}^2_{\boldsymbol{\rho}h}\right) \tag{30}$$

*where $\longrightarrow_{\mathcal{L}}$ stands for the convergence in distribution, and with $\mathbf{0}_Q$ a null vector of size $Q$.*

*Suppose in addition that*

*H8: There exists some constants $q^1, \ldots, q^H \in ]0,1[$ such that for any $h = 1, \ldots, H$:*

$$\frac{n^h}{n} \quad \to \quad q^h.$$

*Then*

$$\sqrt{n}\left\{\hat{\boldsymbol{\tau}}_{\boldsymbol{\rho}\pi} - \boldsymbol{\tau}_{\boldsymbol{\rho}}\right\} \quad \longrightarrow_{\mathcal{L}} \quad \mathcal{N}\left(\mathbf{0}_Q, \sum_{h=1}^{H} \frac{\boldsymbol{\Sigma}^2_{\boldsymbol{\rho}h}}{p_h}\right). \tag{31}$$

If the number of strata $H$ remains fixed as $n \to \infty$, the assumption (H8) is automatically satisfied for equal allocation $n^h = n/H$, proportional allocation $n^h = n\frac{A^h}{A}$, or Neyman allocation, for example.

By the mean-square consistency of $\hat{V}_{HT}(\rho\pi)$ and $\hat{V}_{YG}(\rho\pi)$ (see equations 19 and 20), we obtain under the assumptions of Theorem 1 that

$$\frac{\hat{V}_{HT}(\rho\pi)}{V_p(\rho\pi)} \to_{Pr} 1 \quad \text{and} \quad \frac{\hat{V}_{YG}(\rho\pi)}{V_p(\rho\pi)} \to_{Pr} 1, \tag{32}$$

where $\to_{Pr}$ stands for the convergence in probability. Therefore, an approximate two-sided $100(1-2\alpha)\%$ confidence interval for $\tau_\rho$ is given by

$$\left[\hat{\tau}_{\rho\pi} \pm u_{1-\alpha}\left\{\hat{V}_{HT}(\rho\pi)\right\}^{0.5}\right] \quad \text{or} \quad \left[\hat{\tau}_{\rho\pi} \pm u_{1-\alpha}\left\{\hat{V}_{YG}(\rho\pi)\right\}^{0.5}\right] \tag{33}$$

with $u_{1-\alpha}$ the quantile of order $1-\alpha$ of the standard normal distribution.

**Theorem 2.** *Suppose that $S$ is selected by stratified sampling, with a finite number of strata $H$ and $n^h \geq 2$ inside each stratum $\mathcal{U}^h$. Suppose that assumptions (H1b) and (H4)-(H7) hold. Then the conclusions of Proposition 2 hold true. If in addition assumption (H9) holds, then:*

$$\sqrt{n}A^{-\beta}\left\{\hat{\theta}_\pi - \theta\right\} \quad \longrightarrow_{\mathcal{L}} \quad \mathcal{N}\left(\mathbf{0}_Q, \frac{1}{A^2}\sum_{h=1}^{H} \frac{\{g'(\mu_{\boldsymbol{\rho}})\}^\top \boldsymbol{\Sigma}^2_{\boldsymbol{\rho}h}\{g'(\mu_{\boldsymbol{\rho}})\}}{p_h}\right). \tag{34}$$

# Bibliography


Barabesi, L. (2003), A monte carlo integration approach to horvitz-thompson estimation in replicated environmental designs, *Metron*, 61, pp. 355–374.

Barabesi, L. and Franceschi, S. (2011), Sampling properties of spatial total estimators under tessellation stratified designs, *Environmetrics*, 22, pp. 271–278.

Barabesi, L., Franceschi, S., and Marcheselli, M. (2012), Properties of design-based estimation under stratified spatial sampling with application to canopy coverage estimation, *The Annals of Applied Statistics*, 6, pp 210–228.

Breidt, F. J. and Opsomer, J. D. (2000), Local polynomial regresssion estimators in survey sampling, *Annals of Statistics*, 28, pp. 1026–1053.

Chauvet, G., Bouriaud, O., and Brion, P. (2023), An extension of the weight share method when using a continuous sampling frame, *Survey Methodology*, 49, pp. 139–162.

Cordy, C. B. (1993), An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe, *Statistics and Probability Letters*, 18, pp. 353–362.

Deville, J. C. (1999), Variance estimation for complex statistics and estimators: linearization and residual techniques, *Survey methodology*, 25, pp. 193–204.

Eriksson, M. (1995), Design-based approaches to horizontal-point-sampling, *Forest science*, 41, pp. 890–907.

Gregoire, T. G. and Valentine, H. T. (2007), *Sampling strategies for natural resources and the environment*, CRC Press.

Isaki, C. T. and Fuller, W. A. (1982), Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, 77, pp. 89–96.

Mandallaz, D. (1991), *A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models*, PhD thesis, ETH Zurich.

Mandallaz, D. (2007), *Sampling techniques for forest inventories*. CRC Press.

Stevens, D. L. and Urquhart, N. S. (2000), Response designs and support regions in sampling continuous domains, *Environmetrics*, 11, pp. 13–41.

Vidal, C., Alberdi, I., Hernández, L., and Redmond, J. (2016), *National forest inventories*, Springer.

# Session groupe Statistique bayésienne

# Calibration d'un modèle de pollinisation à l'échelle du paysage par des méthodes de type Approximate Bayesian Computation

Charlotte Baey [1] & Henrik G. Smith [2,3] & Maj Rundlöf [2] & Ola Olsson [2] & Yann Clough [2] & Ullrika Sahlin [2]

[1] *Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France*
*charlotte.baey@univ-lille.fr*

[2] *Lund University, Department of Biology, SE-223 62 Lund, Sweden*
[3] *Lund University, Centre for Environmental and Climate Science, SE-223 62 Lund, Sweden*

**Résumé.** La modélisation des services écosystémiques passe souvent par la construction de modèles mécanistes parfois complexes, dont la calibration peut s'avérer délicate. Dans ce travail, on s'intéresse à un modèle de pollinisation spatialement explicite, appliqué au bourdon terrestre (*Bombus terrestris*). La vraisemblance n'étant pas calculable analytiquement, nous proposons une approche de type Approximate Bayesian Computation (ABC) pour l'estimation des paramètres du modèle. Nous comparons différentes méthodes ABC permettant de prendre en compte la grande dimension des observations. La première étape consiste à définir un ensemble de statistiques résumées sur laquelle les méthodes ABC seront appliquées. Nous considérons ensuite deux stratégies, l'une reposant sur des méthodes de régression pour ajuster les échantillons obtenus selon la loi a posteriori ABC, et l'autre reposant sur des méthodes de type machine learning pour approcher certaines caractéristiques de la loi a posteriori ABC (e.g. moyenne et quantiles). Les résultats obtenus sur données simulées montrent que certains paramètres sont plus faciles à estimer que d'autres, et les approches basées sur des forêts aléatoires s'avèrent plus performantes. L'application aux données réelles montre l'intérêt de l'approche ABC dans le contexte de modèles complexes tout en mettant en évidence les difficultés liées au choix des statistiques résumées et de la méthode.

**Mots-clés.** Statistique bayésienne, méthode ABC, calibration, modèle de pollinisation

**Abstract.** Modelling ecosystem services often requires building mechanistic models which might be complex and which calibration can be challenging. In this work, we are interested in a spatially explicit foraging model for *Bombus terrestris*, accounting to bee distribution in the landscape. The likelihood of the model being intractable, we rely on Approximate Bayesian Computation (ABC) for the estimation of the model parameters. We compare different ABC methods to handle the high dimension of our observations. The first step consists in the definition of a set of summary statistics on which ABC is then applied. Two ABC strategies were then studied. The first one rely on the use of regression adjustment methods, to produce ABC posterior samples. The second one rely on the use of machine learning approaches to approximate key quantities of the ABC posterior distribution (e;g. mean and quantiles). Results from simulated data show that some parameters are easier to calibrate than others, and that approaches based on random forests performed better.

Results on real data show how appealing the methodology can be, even though tuning ABC can be challenging, especially regarding the choice of the summary statistics.

**Keywords.** Bayesian statistics, ABC method, calibration, pollination model

# 1 Introduction

L'étude et l'évaluation des services écosystémiques, que ce soit d'un point de vue écologique ou socio-économique, est un enjeu majeur des politiques publiques. Parmi ces services écosystémiques, on retrouve notamment la pollinisation par les insectes, qui permet entre autres de maintenir la biodiversité des plantes sauvages ainsi que la production d'un grand nombre de plantes cultivées. Dans un contexte de déclin des populations d'insectes pollinisateurs, il est alors crucial de pouvoir estimer avec le plus de précision possible le statut de ces populations, mais aussi de pouvoir prédire l'effet de certaines pratiques agricoles ou paysagères sur ces populations. Ces questions peuvent être explorées à l'aide de modèles écologiques, et dans le cas qui nous intéresse, de modèles spatialement explicites prenant en compte la distribution des insectes dans le paysage.

Les modèles développés dans ce cadre là sont des modèles dits mécanistes, qui sont souvent complexes et hautement non linéaires, produisant des sorties de grande dimension qui peuvent elle-mêmes posséder des structures complexes. La calibration des paramètres de ces modèles peut alors s'avérer délicate. Par exemple, le temps d'exécution d'une instance du modèle peut être élevé, rendant coûteuse toute procédure d'estimation nécessitant des appels répétés au modèle. De plus, ils s'écrivent souvent comme un ensemble de relations hiérarchiques faisant intervenir des variables latentes qui peuvent être également de grande dimension.

Dans ce travail, nous proposons une approche de type *Approximate Bayesian Computation* (ABC), qui permet de contourner les difficultés évoquées plus haut. Nous comparons différentes méthodes de type ABC, reposant sur deux types de stratégies pour la prise en compte de données en grande dimension. La première utilise des méthodes de régression pour ajuster les échantillons obtenus sous la loi a posteriori ABC, et la deuxième utilise des approches de type machine learning pour estimer des quantités clés de la distribution a posteriori ABC. Les méthodes sont comparées sur des données simulées puis appliquées sur un jeu de données réelles.

# 2 Modèle de pollinisation

## 2.1 Modèle mécaniste de type 'Central Place Forager'

Le modèle de pollinisation utilisé (Olsson et Bolin, 2014) est basé sur la théorie du 'central place foraging' qui permet de décrire le comportement d'animaux qui partent à la recherche de nourriture à partir d'un nid ou d'un habitat central. C'est notamment le cas du bourdon

terrestre. A partir de la localisation des nids de bourdons et des sources de nourriture dans le paysage, on peut alors déterminer les ressources qui seront visitées par les bourdons d'un nid donné en fonction de la qualité de la resource et de sa distance au nid.

Tout d'abord, le paysage est divisé en cellules carrées, auxquelles on associe un type d'habitat (ex. : champ de colza, forêt, zone urbaine, ...), puis une qualité de resource et une présence ou absence de nid en fonction du type d'habitat. Pour un bourdon dont le nid est situé sur la cellule $i$ et une resource située sur la cellule $j$, on mesure la différence entre la distance maximale que l'insecte est prêt à parcourir pour une resource de cette qualité, et la distance réelle entre $i$ et $j$ par la quantité suivante : $\Delta_{ij} = \tau_0 \left(1 - \frac{f_0}{f_j}\right) - d_{ij}$, où $\tau_0$ est la distance maximum qu'un bourdon peut parcourir, et $f_0$ la qualité minimale de resource exigée par l'insecte. Puis on définit la qualité d'un nid situé en $i$ par : $s_i = \sum_j \Delta_{ij} \mathbf{1}_{\Delta_{ij}>0}$. Plus il y a de sites avec des ressources de bonne qualité à proximité de la cellule $i$, plus la qualité d'un nid situé en $i$ sera élevée.

En pratique, un bourdon dont le nid est entouré de ressources de bonne qualité aura tendance à rester à proximité de son nid et à peu exploiter les sites situés loin de son nid. On peut alors définir la distance maximale qu'un bourdon est prêt à parcourir à partir de son nid, en fonction de la qualité de celui-ci :

$$\tau_i = \frac{\tau_0}{1 + \exp((\sqrt{s_i} - a)/b)}.$$

Comme pour la définition de $\Delta_{ij}$, on peut définir $\Delta^*_{ij}$ en remplaçant $\tau_0$ par $\tau_i$ :

$$\Delta^*_{ij} = \tau_i \left(1 - \frac{f_0}{f_j}\right) - d_{ij}.$$

L'intensité de visites de bourdons sur le site $j$ est alors définie par :

$$\nu_j(\theta, \mathcal{M}) = \sum_{i=1}^{I} q_i \frac{\Delta^*_{ij}}{\sum_{j=1}^{J} \Delta^*_{ij}},$$

où $q_i$ est une variable binaire indiquant la présence ou l'absence de nid sur la cellule $i$.

## 2.2 Données

On dispose de deux jeux de données provenant du sud de la Suède, sur la surveillance des insectes pollinisateurs dans différents types de paysage présentant un gradient de ressources florales. Les deux études recouvrent 4 années d'observation, et plusieurs observations ont été faites au cours de chaque année d'observation, correspondant à différentes périodes du cycle de vie du bourdon. Au total, on dispose d'un ensemble de 790 observations du nombre de bourdons, relevées dans des transects de longueur fixée, et dans différents types d'habitat : champ de colza, prairie (semi-naturelle), bordure de champ, champ de céréales.

## 2.3 Modèle statistique

On note alors $y_{ijk}, i = 1, \ldots, n, j = 1, \ldots, J, k = 1, \ldots, K$ les observations du nombre de bourdons sur le site $i$, l'année $j$ et à la période $k$.

**Vraisemblance.** On note $\lambda_{ijk}$ l'intensité réelle du taux de visites sur le site $i$, l'année $j$ et la période $k$. Le modèle est alors décrit sous la forme hiérarchique suivante :

$$y_{ijk} \mid \lambda_{ijk}, \theta \sim \mathcal{P}(c_i \cdot \lambda_{ijk}),$$

où $c_i$ est une constante connue permettant de prendre en compte la surface de la zone d'observation et la durée d'observation et $\theta$ représente l'ensemble des paramètres du modèle de pollinisation.

$$\log \lambda_{ijk} = \log \nu_i(\theta, \mathcal{M}_{ijk}) + \beta_1 + \sum_{l=2}^{K} \beta_l \mathbf{1}_{l=k} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2).$$

Les paramètres $\beta_l, l = 1, \ldots, K$ permettent de prendre en compte l'évolution de la taille de la population au cours de la saison.

On note $\psi = (\theta, \beta_1, \ldots, \beta_K, \sigma^2)$ l'ensemble des paramètres du modèle. Le modèle ainsi décrit conduit à une vraisemblance de type Poisson-lognormal, qui n'est pas calculable analytiquement.

**Lois a priori.** Des lois a prioris non informatives ont été choisies en l'absence d'information biologiques disponible sur les processus correspondants. Pour les paramètres $\tau_0$ et $f_0$ dont l'interprétation biologique est plus aisée, l'avis d'experts a été pris en compte.

$$\begin{aligned}
\tau_0 &\sim \mathcal{LN}_{[0,1000]}(\log(1000), 1) \\
f_0 &\sim \mathcal{LN}(\log(0.1), 1) \\
a &\sim \mathcal{U}([100, 1000]) \\
b &\sim \mathcal{U}([100, 1000]) \\
\beta_k &\sim \mathcal{N}(0, 100), \quad k = 1, \ldots, K \\
\sigma^2 &\sim \mathcal{IG}(1, 1)
\end{aligned}$$

# 3 Estimation par la méthode ABC

## 3.1 Choix des statistiques résumées

Une première étape consiste à construire un ensemble de statistiques résumées, afin de procéder à une première réduction de la dimension. Le choix suivant a été fait, en concertation avec les biologistes : intervalle interquartile et nombre de 0 observés par site, par

période et par an, tout types d'habitat confondus, et intervalle interquartile et nombre de 0 observés par type d'habitat, par période et par an, tous sites confondus. Ce choix a permis de passer de 790 observations à 404 statistiques résumées.

Puis, afin de définir la distance entre les statistiques résumées observées et celles qui seront calculées lors des procédures ABC, on utilise un noyau d'Epanechnikov, dont l'échelle est réglée de sorte à ne conserver qu'une proportion $\epsilon$ des valeurs simulées qui sont les plus proches des valeurs observées.

## 3.2 Méthodes pour l'approximation de la loi a posteriori ABC

Le premier point de vue adopté s'attache à construire des échantillons suivant la loi a posteriori ABC, en utilisant des méthodes de régression pour ajuster les valeurs de paramètres associées à des statistiques résumées situées au voisinage des statistiques résumées observées. L'idée consiste à construire tout d'abord un modèle de régression des paramètres du modèle sur les statistiques résumées (Blum et François, 2010) :

$$\psi_i^{(m)} = m_i\left(s^{(m)}\right) + \sigma_i\left(s^{(m)}\right)\varepsilon_{im}, \quad i = 1, \ldots, p \tag{1}$$

où les $\varepsilon_{im}$ sont des variables iid centrées, et où la fonction $\sigma_i$ permet de prendre en compte une éventuelle hétéroscédasticité. Une fois le modèle de régression estimé, on peut construire des valeurs ajustées pour les paramètres retenus de la façon suivante :

$$\psi_i^{*(m)} = \hat{m}_i(s_{\text{obs}}) + (\psi_i^{(m)} - \hat{m}_i(s^{(m)}))\,\frac{\hat{\sigma}_i(s_{\text{obs}})}{\hat{\sigma}_i(s^{(m)})}, \quad i = 1, \ldots, p.$$

Quatre approches ont été comparées : i) $m_i$ linéaire et erreurs hétéroscédastiques, ii) $m_i$ non linéaire et erreurs hétéroscédastiques, iii) approche en deux étapes où (1) est appliquée une première fois pour obtenir des échantillons ajustés à partir desquels on estime le support de la loi a posteriori ABC, puis on ré-applique (1) sur les échantillons ajustés appartenant à ce support, iv) $m_i$ non linéaire et erreurs homoscédastiques, estimé par un modèle de forêt aléatoire.

## 3.3 Méthodes pour l'approximation de caractéristiques de la loi a posteriori ABC

Le second point de vue consiste à approcher certaines quantités d'intérêt de la loi a posteriori ABC, sans chercher à reconstruire toute la distribution. On s'est intéressé à l'estimation de la moyenne, de la médiane et des quantiles d'ordre 0.025 et 0.975 de la loi a posteriori ABC.

Deux approches ont été testées : i) régression (classique et quantile) par forêt aléatoire (voir Raynal et al. 2018), ii) régression par méthode de boosting. Dans le premier cas, on a comparé les forêts aléatoires pondérées ou sans poids, en utilisant les poids donnés par le noyau d'Epanechnikov, et dans le deuxième cas on a comparé deux fonctions de perte, la perte $L_1$ et la perte $L_2$.

# 4 Résultats

## 4.1 Données simulées

Les méthodes ont d'abord été comparées sur un jeu de données simulées, selon la procédure suivante : (i) $M = 100\ 000$ valeurs de paramètres ont été échantillonnées à partir des lois a priori, et $M$ jeux de données simulées ont été construits à l'aide de ces valeurs de paramètres, (ii) 100 jeux de données ont été choisis aléatoirement parmi les 100 000, pour jouer le rôle de jeux de données de référence, et (iii) les différentes méthodes ont été appliquées sur chacun des 100 jeux de données de référence pour obtenir soit des échantillons selon la loi a posteriori ABC soit des quantités unidimensionnelles de cette loi a posteriori, en utilisant les 999 900 jeux de données restants.

Les performances des méthodes ont été comparées à l'aide de l'erreur relative absolue (RAE) entre la médiane a posteriori et la vraie valeur du paramètre (voir Figure 1), et à l'aide de la couverture empirique des intervalles de crédibilité obtenus.

Les méthodes basées sur les forêts aléatoires sont associées aux valeurs les plus faibles de l'erreur relative absolue, quel que soit le paramètre considéré. La méthode de régression adaptative fonctionne en général mieux que les méthodes non adaptatives, et les méthodes basées sur une régression linéaire locale produisent des lois a posteriori aux supports trop larges, ne respectant pas en particulier le support de la loi a priori. En pratique, ces approches n'ont parfois pas abouti à cause de problèmes de convergence.

## 4.2 Données réelles

Les résultats obtenus sur les données réelles diffèrent selon le paramètre considéré. D'un côté, pour les paramètres du modèle mécaniste de pollinisation, i.e. $\theta = (\tau_0, f_0, a, b)$, les méthodes de type rejet ou utilisant le gradient boosting produisent des intervalles identiques à ceux que l'on obtiendrait avec la loi a priori alors que les autres approches produisent des intervalles sensiblement différents. D'un autre côté, pour les paramètres du modèle d'observation, i.e. $\omega = (\beta_1, \beta_2, \beta_3, \sigma^2)$, toutes les approches produisent des intervalles de crédibilité qui diffèrent de ceux obtenus avec la loi a priori. Pour les méthodes basées sur des forêts aléatoires, les résultats obtenus pour le paramètre $\sigma^2$ indiquent une sur-estimation de ce paramètre, ce qui pourrait être dû à une sous-estimation de l'intensité sous-jacente du processus de Poisson.

# 5 Conclusion

L'objectif de ce travail était de proposer une méthodologie pour la calibration de modèles complexes pour lesquels la vraisemblance n'est pas calculable analytiquement. Différentes méthodes de type ABC ont été comparées, et les avantages et inconvénients de chacune d'entre elles ont été discutés. Il ressort de cette étude que les approches basées sur les forêts aléatoires, se concentrant sur l'approximation de quantités unidimensionnelles de la

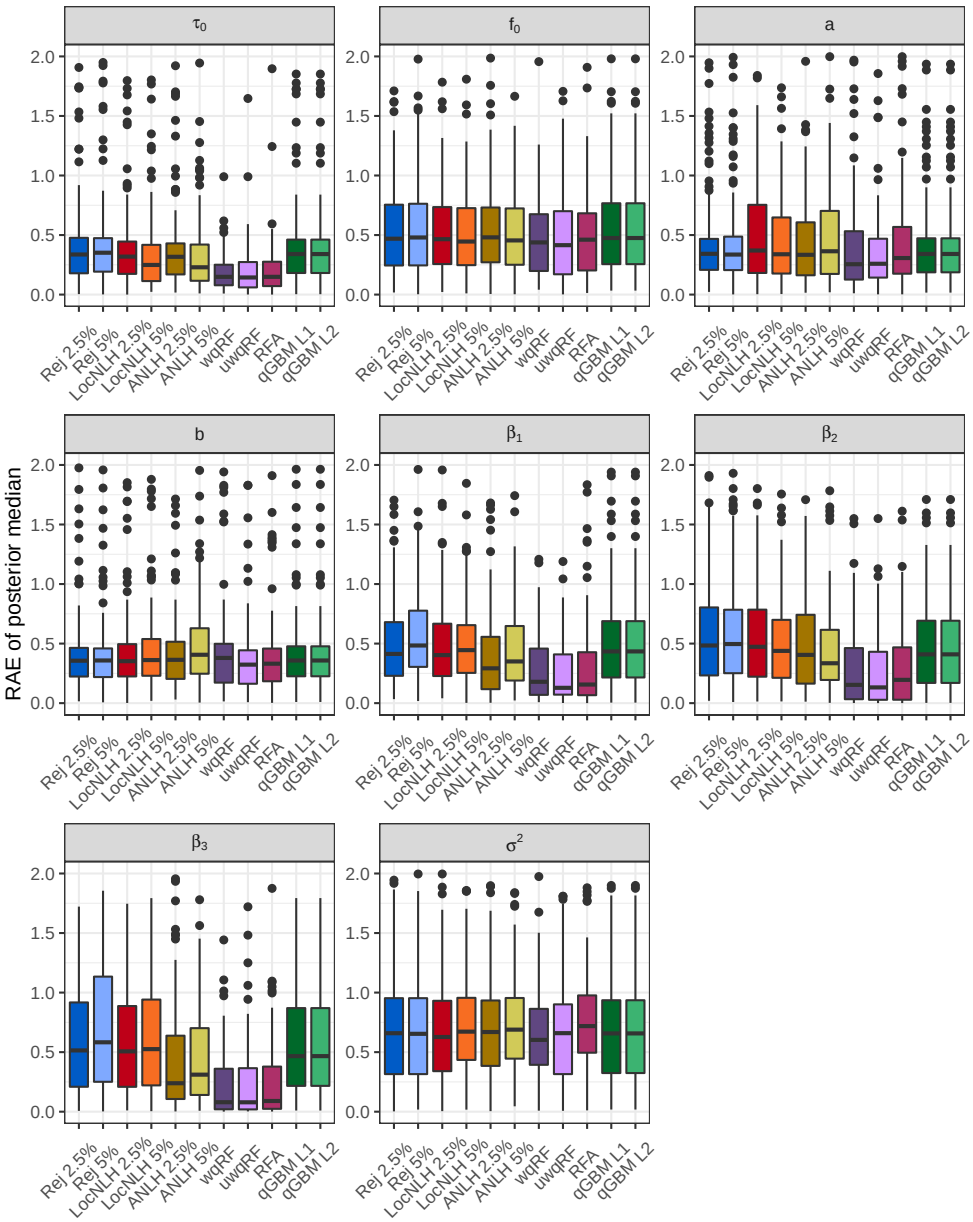



Figure 1: Erreur relative absolue en utilisant la médiane a posteriori

loi a posteriori ont de meilleures performances sur notre modèle. Elles sont particulièrement simples à mettre en place.

En revanche, certains paramètres restent difficiles à estimer, quelle que soit la méthode utilisée, ce qui souligne les difficultés auxquelles on peut faire face dans ce type de contexte. Plusieurs situations peuvent être à l'origine de cette difficulté d'estimation : la loi a priori est mal choisie, le modèle n'est pas identifiable, les données ne sont pas suffisamment informatives, le choix des statistiques résumées n'est pas adapté. Chacune de ces pistes peut être poursuivie afin d'améliorer le modèle, ou le recueil des données.

# Bibliographie

C Baey, H. G; Smith, M. Rundlöf, O. Olsson, Y. Clough, U. Sahlin (2023), Calibration of a bumble bee foraging model using Approximate Bayesian Computation, *Ecological Modelling*, 477: 110251.

M. G. Blum and O. François (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and computing*, 20(1):63–73.

O. Olsson and A. Bolin (2014). A model for habitat selection and species distribution derived from central place foraging theory. *Oecologia*, 175(2):537–548.

L. Raynal, J.-M. Marin, P. Pudlo, M. Ribatet, C. P. Robert, and A. Estoup (2018), ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

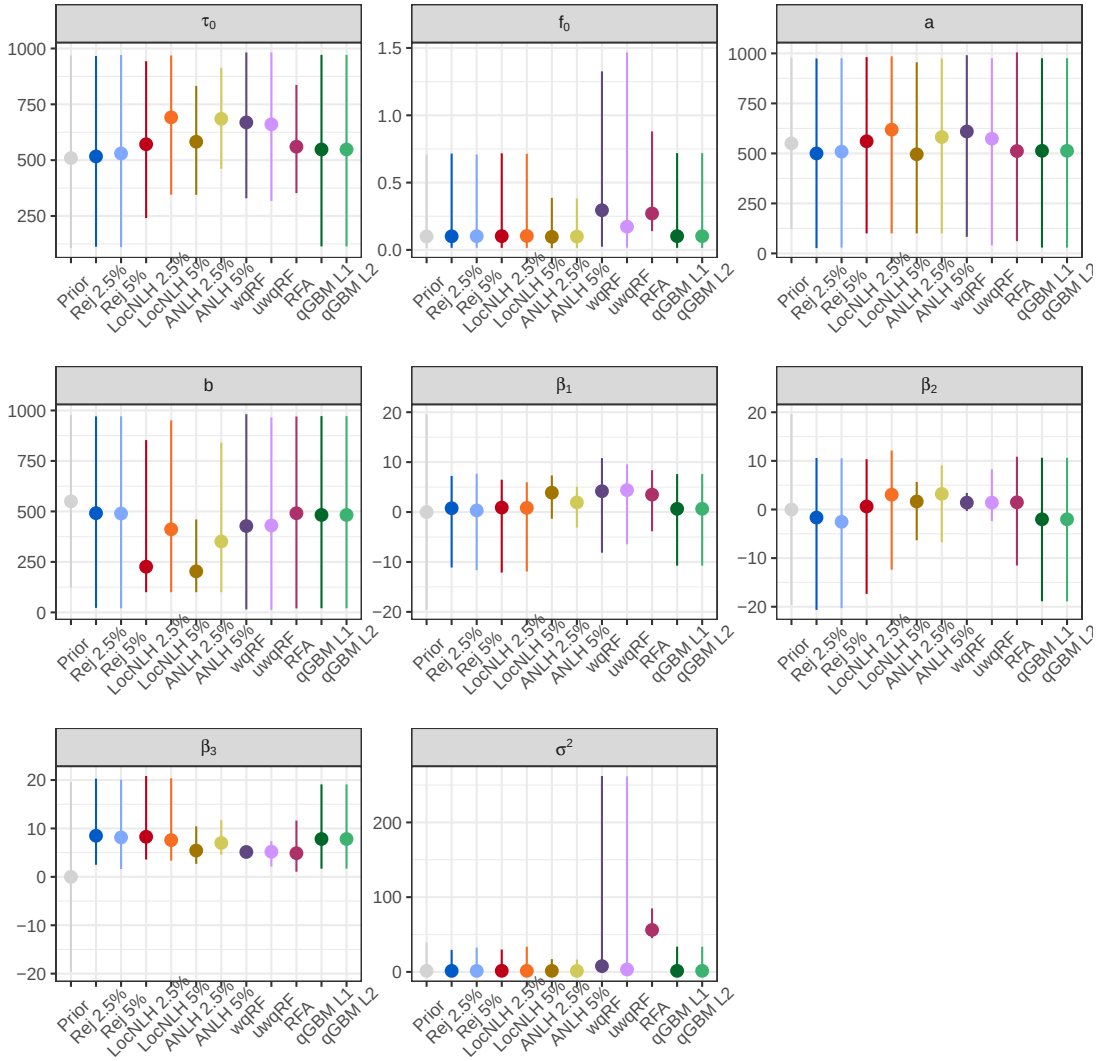



Figure 2: Intervalle de crédibilité à 95% sur les données réelles.

# Développement d'un modèle stochastique de surproduction en vue de gérer les captures accidentelles d'espèces protégées

Fanny Ouzoulias [1], Nicolas Bousquet [2], Mathieu Genu [3], Anita Gilles [4], Jérôme Spitz [5] & Matthieu Authier [6]

[1] *Laboratoire de Biologie des Organismes et Ecosystèmes Aquatiques (BOREA), UMR 8067 - MNHN, CNRS, IRD, SU, UCN, UA, 75005, Paris, France. fanny.ouzoulias@mnhn.fr*
[2] *Laboratoire Probabilités, Statistiques et Modélisation, UMR 8001 CNRS, Sorbonne Université, France. nicolas.bousquet@sorbonne-universite.fr*
[3] *Observatoire Pelagis, UAR 3462 CNRS - La Rochelle University, France. mgenu@univ-lr.fr*
[4] *Institute for Terrestrial and Aquatic Wildlife Research, University of Veterinary Medicine Hannover, Foundation, Büsum, Germany. anita.gilles@tiho-hannover.de*
[5] *entre d'Etudes Biologiques de Chizé, UMR 7372 CNRS-LRUniv, 79360 Villiers en Bois, France. jspitz@univ-lr.fr*
[6] *Observatoire Pelagis, UAR 3462 CNRS - La Rochelle University, France. mauthier@univ-lr.fr*

**Résumé.** La gestion des prises accidentelles d'espèces protégées dans les engins de pêche est un enjeu majeur pour atteindre les objectifs de la Stratégie de l'Union Européenne en faveur de la biodiversité à horizon 2030. Les statistiques des prélèvements réalisés ne sont, en général, pas fiables car non-systématiques, ou issues déchantillons non-représentatifs. Les données de prélèvements concernant les espèces protégées sont donc très parcellaires et biaisées, ce qui complique l'évaluation de la soutenabilité des activités humaines. Toutes les espèces de cétacés sont protégées au niveau national et européen ; néanmoins celles-ci sont aussi impactées par les captures accidentelles. C'est le cas du dauphin commun (*Delphinus delphis*) dans le Golfe de Gascogne, dont les manquements en matière de protection valent aujourd'hui à la France une mise en demeure de la Commission Européenne et à l'Etat un arrêté de fermeture des pêcheries à risque pendant un mois durant l'hiver 2024 ordonnée par le Conseil d'Etat.

Un outil important pour gérer les impacts des pêcheries est le calcul de points de référence limites, aussi appelé seuils de prélèvements au-delà desquels la viabilité à long terme des populations impactées n'est plus garantie. Le calcul de tels seuils repose sur une méthodologie mise en place par la Commission Baleinière Internationale (CBI) au cours des années 1990 et qui reposent sur la simulation numérique de populations virtuelles soumises à des prélèvements dont l'ampleur est déterminée par une règle de gestion. Ces règles de gestion pour calculer les seuils sont évaluées dans un panel de scénarios pour évaluer la robustesse de ceux-ci à divers cas de figures dont des biais dans les données (sous-estimation des prélèvements par sous-déclaration, etc.). Deux règles sont actuellement utilisées : le PBR ("Potentiel Biological Removal") tel que défini dans la loi états-unienne sur la protection des mammifères marins ; et le RLA ("Removals Limit Algorithm") inspiré des travaux de la CBI.

Nous avons développé un modèle stochastique dit de "sur-production", modèle couramment utilisé en halieutique pour calculer des points de références ou des quotas pour des espèces exploitées. Ce modèle paramétrique reste simple (4 paramètres à estimer) et est informé par les prélèvements et l'abondance estimées d'une espèce pour calculer un taux de prélèvement (en pourcentage de l'abondance) compatible avec la viabilité à long-terme d'une population. Néanmoins, il fait une hypothèse restrictive de stationnarité des prélèvements qui n'est pas réaliste : la gestion doit précisément amener à des taux non-stationnaires puisqu'un objectif de restauration de la biodiversité est de minimiser au cours de temps les captures accidentelles. Nous proposons une approche de vraisemblance pondérée pour s'accommoder de cette non-stationnarité. Au travers d'une étude de simulations, nous montrons comment notre modèle conduit à une règle de gestion compatible avec les ambitions actuelles de conservation, notamment celle de minimiser l'impact des activités humaines sur les espèces protégées, dauphins inclus.

**Mots-clés.** évaluation de stratégie de gestion, `Stan`, captures accidentelles, cétacés, Bayesien

**Abstract.** Managing human activities, which can result in additional mortality on many marine Protected, Endangered or Threatened Species (PETS) is key to reach the ambitious set out by the EU 2030 Biodiversity Strategy. By-catch, the undesirable and non-intentional catch of non-target species in marine fisheries, is one of the main causes of mortality of marine mammals (which are often PETS) worldwide. Data on anthropogenic removals (including by-catch) of PETS (including marine mammals) are often unreliable because of, among others, inadequate sampling design, lack of enforcement, non-representative samples and underreporting (due for example to social desirability bias). All cetacean species benefit from some legal protection whether at the national or international level. The common dolphin (*Delphinus delphis*) in the Bay of Biscay epitomized the current challenges : the failure to enforce its strict protection earned (i) France an infringement procedure from the European Commission ; and (ii) the French Government the ordnance from the highest administrative court ('The Conseil d'Etat') of a one month spatio-temporal closure of all high-risk fisheries operating in the Bay of Biscay in the winter 2024.

Managing by-catch hinges on the computation of so-called biological reference points, also known as removals limits/thresholds : these represent an upper limit to the number of animals that can be removed from a population without compromising the long-term viability of said population with unacceptably high probability. Methods to compute removals limits for cetaceans originate from scientific work carried out in the 1990s by the International Whaling Commission (IWC) whereby computer simulations are harnessed to investigate the likelihood outcomes of different management schemes. The framework outlined by the IWC used so-called 'harvest control rule' (or just control rule) that take data from current monitoring as inputs to output a threshold (or a quota in case of a commercial species). Importantly, the framework allows to assess the effect of knowledge gaps and data biases to devise control rules that are robust against these. Two rules are commonly in use : the Potential Biological Removal (PBR) from the US Marine Mammal Protection Act ; and the Removals Limit Algo-

rithm, a child of the Catch Limit Algorithm devised to set quotas on the hunt of large whales.

We developed a stochastic surplus production model, a kind of parameter-lean model common in fishery sciences, and proposed a new control rule derived from this 4-parameters model. The model assumes (i) a simple proportional relationship between true abundance and removals, and (ii) stationarity (time-invariance) in removal rate. This assumption is untenable if management is to be effective as the very purpose managing anthropogenic removals of PETS is to minimize them. To preserve parameter-leaness, we resorted to a weighted-likehood approach for estimation (in a Bayesian framework) with time-dependent weights chosen such that older removal data are progressively and smoothly down-weighted. Using simulations, we benchmarked our new control rule relying on the stochastic surplus production model in a case study which revealed the competitiveness of the new rule to meet current conservation policy desiderata such as minimizing removals over time.

**Keywords.** management strategy evaluation, `Stan`, by-catch, cetaceans, Bayesian

# 1 Texte long

## Introduction

Human activities in the oceans are increasing and can result in additional mortality on many marine Protected, Endangered or Threatened Species (PETS). By-catch, the undesirable and non-intentional catch of non-target species in marine fisheries, is one of the main causes of mortality of marine mammals (which are often PETS) worldwide. When quantitative conservation objectives and management goals are clearly defined, computer-based procedures can be used to explore likely population dynamics under different management scenarios and estimate the levels of anthropogenic removals, including by-catch, that marine mammal populations may withstand. Two control rules for setting removals limits are the Potential Biological Removal (PBR; Wade 1998) established under the US Marine Mammal Protection Act and the Removals Limit Algorithm (RLA; Cooke 1999) inspired from the Catch Limit Algorithm developed under the Revised Management Procedure of the International Whaling Commission (IWC). Both rules were tested and developped in procedures originally labeled 'simulation trials' and nowadays called Management Strategy Evaluations (MSE).

A management strategy is an agreed-upon set of rules for determining thresholds beyond which a conservation objective runs the risk of not being met with unacceptably high probability. This strategy defines management objectives in the form of thresholds that managers can monitor from available data, with the management objectives that these thresholds are not exceeded. MSE needs generative models that can generate (synthetic) data that are similar to observed, and crucially, currently available data. These models need to be more than simple curve-fitting devices and should be infused with ecological realism to reproduce and simulate the dynamics of an ecological system such as a population subjected to anthropogenic removals on top of natural processes (*e.g.* density dependence). Scientists can then evaluate the performance of management actions in 'what-if', or counterfactual, scenarios to set efficient management objectives. Importantly, the latter will be gauged against observable and available data (*e.g.* abundance and by-catch estimates, along with their uncertainties) only and not from unknown quantities (*e.g.* true abundance). Uncertainties in the underlying model and potential biases and uncertainty in the observed data must be considered in order to ensure robust management.

MSE requires in practice several components, including:
(1) one or several unambiguous quantitative conservation objective;
(2) a data simulator (or operating model) to emulate population dynamics and the effects of anthropogenic activities on this population;
(3) a control rule, whose computation accounts for the expected quantity and quality of observable data; to set a removals limit beyond which the impact of human activities runs the risk of failing the conservation objective(s) in (1); and
(4) performance metrics, necessarily context-dependent and policy-relevant, reflecting the trade-off between the potentially many conservation objectives defined previously.
For each management strategy, population dynamics are simulated, human activities have impacts, a control rule is applied: performance metrics are monitored and ultimately assessed

with respect to the conservation objective. Items (1) and (4) should be agreed upon by all stakeholders or taken from national or international law. Items (2) and (3) are under the remit of scientists, whose task is to test a large panel of realistic scenarios to buffer the management strategy against uncertainties and potential biases in the available data. MSE is computer intensive and needs tuning via simulations. Running a large number of simulations has become mundane yet coding an adequate data simulator may present a daunting task. To minimize duplication of effort and to enhance reproducibility we wrote the `RLA` package for statistical software `R` for ecologists and managers (Genu et al. 2021). The package allows to carry MSE with the two control rules: Potential Biological Removal (PBR) and the Removals Limit Algorithm (RLA). We added a new one: Anthropogenic Removals Threshold or, simply, ART.

## Material and Methods

### Notation

Notations are summarized in Table 1. Let $\log \mathcal{N}(\text{location}, \text{scale})$ denotes the log-normal distribution of parameters location and scale. The $\widehat{\ }$ notation flags a point estimate of a parameter (*e.g.* a quantile from a posterior distribution).

| Name | Type | Meaning |
|---|---|---|
| $K$ | Integer | Carrying capacity (same unit as $N_t$, $N_t^{\text{obs}}$ or $R_t$) |
| $N_t$ | Integer | True abundance (in number of individuals) at time $t$ |
| $N_t^{\text{obs}}$ | Integer | Observed abundance (in number of individuals) at time $t$ |
| $\text{cv}_t$ | Positive real | Coefficient of variation associated with $N_t^{\text{obs}}$ |
| $R_t$ | Integer | Removals (in number of individuals) at time $t$ |
| $D_t$ | Positive real | Depletion at time $t$: ratio of $N_t$ over $K$ |
| $\rho$ | Positive real | Removal rate |
| $r^*$ | Positive real | Population growth rate at the MNPL |
| $r$ | Positive real | Current population growth rate |
| MNP | Positive real | Maximum Net Productivity:<br>the maximum possible *per capita* rate of increase per year |
| MNPL | Proportion | Maximum Net Productivity Level |
| z | Positive real | Shape parameter of the Generalized Logistic Population Growth model |
| $r_{\max}$ | Positive real | Maximum theoretical or estimated productivity rate; related to MNP |
| $F_{\text{R}}$ | Proportion | Recovery factor |
| $\text{N}_{\min}$ | Integer | Minimum population estimate (Wade 1998) |
| IPL | Proportion | Internal Protection Level; a fraction of $K$ |
| $w_t$ | Positive real | weight for the likelihood (Eq. 14) |
| $\text{cv}_\sigma$ | Positive real | Coefficient of variation associated with environmental stochasticity |
| $\varepsilon_t$, $\sigma$ | Positive real | Environmental stochasticity |

Table 1: **Notations**.

**Potential Biological Removal**

The calculation of PBR is model-free:

$$\text{PBR} = N_{\min} \frac{1}{2} r_{\max} F_{\text{R}},$$

$N_{\min}$ is an estimate of minimum population size, $\frac{1}{2} r_{\max}$ one half of the maximum theoretical net productivity rate, and $F_{\text{R}}$ a recovery factor between 0.1 and 1 (Wade 1998). The computation of PBR does not require data on removals.

**Removals Limit Algorithm**

The other harvest control rule currently in use is the Removals Limit Algorithm (RLA). Its computation requires both a time-series of abundance/biomass estimates (whereas PBR only requires one such estimate) and a time-series of removals (whereas PBR requires none). RLA is a variant of the Catch Limit Algorithm for baleen whales (Cooke 1999):

$$N_{t+1} = N_t + r N_t \left(1 - \left(\frac{N_t}{N_0}\right)^2\right) - R_t, \tag{1}$$

where $N_t$ and $R_t$ are respectively the abundance/biomass and removals at time $t$. The computation of the RLA control rule for setting a removals limit (as a fraction of the best available abundance estimate) is:

$$\text{removals limit} = r \times \max\left(0, D_T - \text{IPL}\right), \tag{2}$$

where $T$ is the current year, $D_T$ current depletion (that is, $D_T = \frac{N_T}{K}$, $K$ being the carrying capacity) and IPL (Internal Protection Level) the depletion level below which the limit is set to 0. Both $r$ and $D_T$ are estimated from the model defined by Eq. 1 in a Bayesian framework and removals limit is computed from the joint posterior distribution of ($r$, $D_T$). A point estimate is used in practice by selecting a quantile of the posterior distribution to account for uncertainty.

**Candidate Control Rules**

The PBR control rule takes a value for $r_{\max}$ as an input while the RLA control rule uses a posterior distribution of $r$ (from the model defined by Eq. 1). For most species, both $r_{\max}$ or $r$ are unknown: a default value can be used for PBR, or $r$ needs to be estimated from a prior and data. This knowledge gap may be exploited to argue against the use of either of these rules using uncertainty distortion strategies (Schweder 2000; Rayner 2012). Devising a new rule to set a removals limit that does not directly hinge on knowledge of this input is desirable to (i) avoid any strategic mis-representation of uncertainty (see Rayner 2012); and (ii) diversify options for discussions during the policy process. We developed candidate control rules based on the same data requirements as the RLA, namely a time-series of removals and at least one estimate of abundance that are fed into a statistical model. The statistical model is, however, different in **how** it incorporates the removals data. In Eq. 1, removals are treated as a known covariate. Below, we develop a stochastic model for removals directly.

**Development of a stochastic Surplus Production Model**

Operating models for PETS are often based on Surplus Production Models (SPM) which are standard models of population dynamics in situations of strong uncertainty and low information. SPMs seek to encompass important population processes governing the dynamics of abundance change over time:

$$\text{next abundance} = \text{previous abundance} + \text{recruitment} + \text{growth} - \text{natural mortality} - \text{anthropogenic removals}$$

Density-dependence is taken into account (Pella & Tomlison, 1969), assuming a first-order Markovian process on abundance:

$$N_{t+1} = N_t + r^* \left(\frac{z+1}{z}\right) N_t \left(1 - \left(\frac{N_t}{K}\right)^z\right) - R_t \tag{3}$$

Setting $z = 1$ gives the Schaefer model: $N_t$ and $R_t$ are respectively the abundance/biomass and removals at time $t$, $K$ the carrying capacity and $r^*$ the growth rate at the Maximum Net Productivity Level (MNPL; $r^*$ is also known as the Maximum Sustainable Yield Rate). Incorporating environmental variability (the so-called process noise $\varepsilon_t$) in a multiplicative way in Eq.3 yields:

$$N_{t+1} = \left\{ N_t + r^* \left(\frac{z+1}{z}\right) N_t \left(1 - \left(\frac{N_t}{K}\right)^z\right) - R_t \right\} \varepsilon_t. \tag{4}$$

where $\varepsilon_t$ is assumed to be unbiased ($\mathbb{E}[\varepsilon_t] = 1$) and homoskedastic ($\mathbb{V}[\varepsilon_t] = \sigma^2$).Assuming a simple relationship between removals $R_t$ and abundance $N_t$:

$$R_t = \rho N_t \tag{5}$$

where $\rho \in ]0, 1[$ is a time-invariant removal rate (Bousquet et al. 2008; Bordet & Rivest 2014), the removal process becomes:

$$R_{t+1} = \left\{ R_t + \frac{z+1}{z} r^* R_t \left(1 - \left(\frac{R_t}{K\rho}\right)^z\right) - \rho R_t) \right\} \varepsilon_t \tag{6}$$

The set of parameters in Eq. 6 is $\theta = \{K, \sigma, \rho, r^*\}$. Parameter $z$ is usually fixed rather than estimated. Setting $z = 2.39$ corresponds to a MNPL of 60% of $K$ as customarily assumed for marine mammals. Eq. 5 is a simplifying assumption that allows to link the abundance and removal processes. Stochasticity is introduced in Eq. 6 for estimating a removal rate from data, meaning that removals are used as an index of abundance.

**Reparametrization**

Following Bordet & Rivest (2014), let

$$Z_t = \frac{N_t}{K} \left(\frac{r^*(z+1)}{z - \rho z + r^*(z+1)}\right)^{\frac{1}{z}}$$

which is well defined for $z - \rho z + r^*(1+z) > 0$, that is

$$\rho \ < 1 + r^* \tfrac{z+1}{z}.$$

With $z > 0$ this assumption is not restrictive since $0 < \rho < 1$. Eq. 4 and 6 can be re-arranged:

$$\begin{aligned} Z_{t+1} &= \left(1 - \rho + r^* \frac{z+1}{z}\right) Z_t \left(1 - Z_t^z\right) \varepsilon_t, \\ R_{t+1} &= \left(1 - \rho + r^* \frac{z+1}{z}\right) R_t \left(1 - \{D(\theta) R_t\}^z\right) \varepsilon_t \end{aligned}$$

with $Z_t = D(\theta) R_t$ and

$$D(\theta) \ = \ \frac{1}{K\rho} \left(\frac{r^*(z+1)}{z - \rho z + r^*(z+1)}\right)^{\frac{1}{z}}.$$

Positive removals imply $R_t \leq \frac{1}{D(\theta)}$. To simplify notations and to adopt a more conventional Markovian writing:

$$R_t \ = \ g(R_{t-1}, \theta) \varepsilon_t \tag{7}$$

where $g$, which is neither linear nor log-linear, is:

$$g(R_{t-1}, \theta) = \left(1 - \rho + r^* \frac{z+1}{z}\right) R_{t-1} \left(1 - \{D(\theta) R_{t-1}\}^z\right). \tag{8}$$

A sequence of observed removals $(R_0, \ldots, R_T)$ informs on $\theta$ via a likelihood function:

$$\ell(R_0, \ldots, R_T | \theta) \ = \ \ell(\{R_t\} | \theta) = \prod_{t=1}^{T} f(R_t | R_{t-1}, \theta) \tag{9}$$

where each conditional density function $f(R_t | R_{t-1}, \theta)$ is determined by a choice on the distribution of $\varepsilon_t$. Although the considered quantities are discrete and bounded in our setting, a log-normal assumption is a customary choice to model environmental stochasticity:

$$\varepsilon_t \ \sim \ \log \mathcal{N}\left(-\frac{\sigma^2}{2}, \sigma\right),$$

which implies both $\mathbb{E}\left[\varepsilon_t | \mathcal{F}_{t-1}\right] = 1$ and $\mathbb{V}\left[\varepsilon_t | \mathcal{F}_{t-1}\right] = \sigma^2$. Accordingly, given Eq. 7, the conditional density of $R_t$ in Eq. 9 becomes, for $t > 1$:

$$f(R_t | R_{t-1}, \theta) = \frac{1}{\sqrt{2\pi}\sigma R_t} \exp\left(-\frac{1}{\sigma^2} \exp\left\{\log R_t - \log g(R_{t-1}, \theta) + \frac{\sigma^2}{2}\right\}^2\right). \tag{10}$$

where $g(R_{t-1}, \theta)$ is given by Eq. 8.
One of the parameters in $\theta$, $K$ needs data on absolute abundance. Denote $\mathcal{I}$ the years for which an abundance estimate $N_t^{\text{obs}}$, observed with noise, is available. We denote $\mathcal{J}$ the years for which removals are observed, with $\mathcal{I} \subset \mathcal{J}$. The true abundance in year $t \in \mathcal{I}$ is

$$N_t \;=\; \frac{R_t}{\rho}.$$

Assuming a log-normal distribution for observation errors $\epsilon'_t$, $\forall t \in \mathcal{I}$:

$$N_t^{\text{obs}} | R_t, \theta, \tau_t \;=\; N_t \exp(\epsilon'_t) \text{ with } \epsilon'_t \sim \log \mathcal{N}\left(-\frac{\tau_t^2}{2}, \tau_t\right),$$

where $\tau_t = \sqrt{\log\left(1 + \text{cv}_t^2\right)}$ and $\text{cv}_t$ is the coefficient of variation associated with the estimated abundance $N_t^{\text{obs}}$. The observation model for each observed abundance datum is:

$$N_t^{\text{obs}} | R_t, \theta, \tau_t \;\sim\; \log \mathcal{N}\left(\log R_t - \log \rho - \frac{\tau_t^2}{2}, \tau_t\right) \tag{11}$$

The joint likelihood of the observed abundances and removals data is

$$\ell_\theta\left(\left\{N_t^{\text{obs}}\right\}_{t\in\mathcal{I}}, \{R_t\}_{t\in\mathcal{J}}\right) = \ell\left(\left\{N_t^{\text{obs}}\right\}_{t\in\mathcal{I}} | \{R_t\}_{t\in\mathcal{J}}, \theta\right) \times \prod_{t\in\mathcal{J}} \ell\left(\{R_t\}, \theta\right) \tag{12}$$

where $\ell\left(\{R_t\}, \theta\right)$ is given by Eq. 9. Under the assumption that abundances are observed independently from removals and given Eq. (11), one has:

$$\ell\left(\left\{N_t^{\text{obs}}\right\}_{t\in\mathcal{I}} | \{R_t\}_{t\in\mathcal{J}}, \theta\right) = \prod_{t\in\mathcal{I}} \frac{1}{\sqrt{2\pi} N_t^{\text{obs}} \tau_t} \exp\left(-\left(\log\left(\frac{\rho N_t^{\text{obs}}}{R_t}\right) + \frac{\tau_t^2}{2}\right)^2 \frac{1}{2\tau_t^2}\right). \tag{13}$$

We wrote the joint likelihood (Eq. 12, Ouzoulias et al. 2024) in programming language `Stan` (Carpenter et al. 2017).

**Initial conditions**
For practical reasons, the initial depletion $D_0$ at the start of the time-series of removals, instead of $K$, is estimated: $D_0 = \frac{N_0}{K}$. Initial abundance $N_0$ is typically unknown for PETS and the first abundance estimate available may not even match the start of the observed removals time-series. In that case, information on initial depletion $D_0$ may be elicited from expert knowledge or historical data, and given a prior distribution: it may be easier to elicit a prior on depletion (a quantity expected to be bounded between 0 and 1) than on $K$ directly. In practice, $K$ is deduced from $D_0$ and the first observed abundance estimate that is available. The set of parameters to estimate is now: $\theta = \{r^*, \sigma, \rho, D_0\}$.

**Anthropogenic Removals Threshold (ART)**
The stochastic SPM assumes a crude proportionality between removals and abundance, and stationarity in $\rho$, the removal rate, which is at odds with the very purpose of managing

removals. If management is meant to be effective, it will take action to precisely change the level of removals. By definition, management aims at changing $\rho$ over time so the model is clearly wrong once management is implemented. Before management is implemented, stationarity is an assumption as there is typically little knowledge or data are too noisy to test it. To obviate this issue while retaining a parsimonious model with 4 parameters, we used a *weighted* likelihood approach to progressively down-weigh data extending the furthest back in time (*i.e.*, to consider that the policy relevance of statistical information provided by each datum in a sample depends on how far back in time the datum was collected). The likelihood $\ell(\{R_t\}|\theta)$ is replaced by

$$\ell^{\boldsymbol{w}(\eta)}(\{R_t\}|\theta) \quad = \quad \prod_{t=1}^{T} \left(f\left(R_t|R_{t-1},\theta\right)\right)^{w_t(\eta)} \tag{14}$$

where the weights $w_t(\eta)$ provide a kernel-based representation of the score function

$$s(\theta) \quad = \nabla_\theta \; \ell(\{R_t\}|\theta).$$

The weights $w_t(\eta)$ should be a bounded differentiable non-negative function of $t$ that may depend on a parameter $\eta$ which can be consistently estimated by $\hat{\eta}$, such that

$$\sup_{R_t} |w(t,\hat{\eta}) - c| \quad \xrightarrow[p]{t\to\infty} \quad 0 \quad \text{almost surely}$$

where $c$ is a positive constant. The following choice (Gaussian kernel):

$$w_t(\eta) \quad = \quad \exp\left(-\frac{(T-t)^2}{2\eta^2}\right)$$

obeys these requirements (with $c = 1$). $\eta$ was fixed instead of estimated so that data older than 50 years contribute less than 0.05 to the likelihood during estimation ($w_t = 0.05$ for $(t - 50)$). This choice ($\eta = 20.4$) is arbitrary but was found to work well in practice. Capitalizing from the stochastic SPM (Eq. 14), we propose two candidate control rules (as a fraction of the best available abundance estimate) which we call *Anthropogenic Removals Threshold* (ART) to emphasize that the quantity derived from these control rules represents a threshold beyond which conservation objectives run a high risk of not being met. The first candidate is simply the posterior mean of the quantity:

$$\text{candidate}_1 = \rho \times F_{\text{R}} \tag{15}$$

where $F_{\text{R}}$ is a recovery factor chosen between 0.1 and 1 (as in the PBR control rule). This rule adapts the historical removal rate and does not directly rely on an estimate of carrying capacity or population growth rate as the RLA control rule (although both a carrying capacity and a population growth rate are in $\theta$). The second candidate takes stock of any decline in abundance to negatively feedback on the removals limit:

$$\text{candidate}_2 = \rho \times F_{\text{R}} \times \min\left(1, \exp\left(\beta\right)\right) \tag{16}$$

where $\beta$ is the slope of a regression line through the abundance estimates (scaled by the first estimate and then log-transformed) and estimated using a weakly-informative prior (namely the 'skeptical' prior of Cook et al. (2011)) that favours the hypothesis of no trend over time. This candidate rule operationalizes the principle on non-deterioration whereby populations or species in need of restoration (that is, that are depleted) should not be allowed to deteriorate further. If no trend or a positive trend in abundance is evidenced, $\text{candidate}_2$ is equivalent to $\text{candidate}_1$. Both $\text{candidate}_1$ and $\text{candidate}_2$ can be computed for the same data necessary to compute RLA, and their posterior mean approximated by the average over a sample from the posterior distribution of $\theta$.

**('Harvest') Control rules**

We tested 4 control rules for managing anthropogenic removals of PETS:

- the PBR rule of Wade (1998) with $N_{\min}$ defined as the 20% quantile of a log-normally distributed abundance estimate $N_T^{\text{obs}}$: $\text{PBR} = N_{\min} \frac{1}{2} r_{\max} F_{\text{R}}$;
- the RLA rule, with $\text{RLA} = N_T^{\text{obs}} \times \text{removals limit} = N_T^{\text{obs}} \times r \times \max(0, D_T - \text{IPL})$;
- the $\text{candidate}_1$ ART, with $\text{ART}_1 = N_T^{\text{obs}} \times \text{candidate}_1 = N_T^{\text{obs}} \times \rho \times F_{\text{R}}$; and
- the $\text{candidate}_2$ ART, with $\text{ART}_2 = N_T^{\text{obs}} \times \text{candidate}_2 = N_T^{\text{obs}} \times \rho \times F_{\text{R}} \times \min(1, \exp(\beta))$.

All rules need tuning. For PBR, $\text{ART}_1$ and $\text{ART}_2$, this process means the testing of different values of $F_{\text{R}}$ to identify the minimum one that allows to reach the conservation objective. For RLA, tuning is achieved by testing different quantiles of the posterior distribution of Eq. 2. That quantile tuning was not carried out with $\text{ART}_1$ or $\text{ART}_2$ stemmed from the typically tight posterior concentration observed when estimating $\rho$ during the development of the stochastic SPM. In contrast, posterior concentration does not occur because the log-normal likelihood assumed for Eq. 1 when estimating removals limit is down-weighted by a fixed factor $\frac{1}{16}$ to limit the speed at which the management procedure responds to feedback (Cooke 1999).

**Simulations**

The operating model used in population dynamics simulations was a stochastic and age-disaggregated version of a generalized logistic model of population dynamics (Genu et al. 2021). Given initial conditions, biological parameters and removals, abundance data are generated at each time step. Life-history parameters of the the Harbour Porpoise (*Phocoena phocoena*) in the North Sea were inputed to the operating model. A hundred (100) simulations were carried out: for each a hypothetical population of harbour porpoises was depleted with unmanaged anthropogenic removals for 50 years before implementing management procedures and specific control rules. A time-series of removals as long as 50 years is unusual in general, but one is available for harbour porpoise in the North Sea. A distribution of initial depletion levels was induced between 30% and 60%. Important biological inputs include the Maximum Net Productivity (MNP) and MNPL, which are usually unknown in most cases. To reflect that uncertainty, a range of plausible values for small cetaceans were considered.

**Scenarios**

We evaluated control rules on three scenarios: a base case scenario whereby unbiased but

noisy data are assumed to be available and collected; and two so-called robustness trials. In the first trial, estimates of abundance have a systematic bias resulting in an overestimation by a factor 2. In the second, removals' estimates were assumed to be biased downward, resulting in an underestimation of true removals by a factor 2. The two robustness trials were found to be the most challenging ones in a previous investigation (Genu et al. 2021). The MSE is summarized on Figure 1. The conservation objective for "long-term viability" was defined as to restore or maintain population size to at least 60% of carrying capacity ($K$) over a time horizon of 50 years with the probability of 0.9.

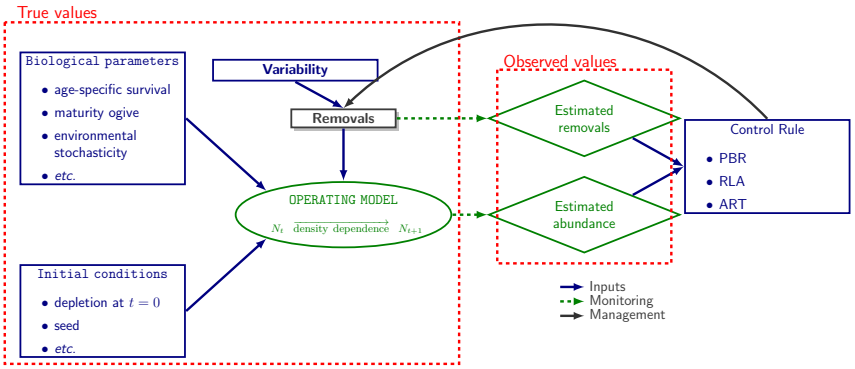



Figure 1: **Simulation workflow.** Schematic representation of the workflow for simulations. Population dynamics ($N$ denotes abundance) are simulated from biological parameters (life-history data, true removal rate etc.). Monitoring allows to collect data but these data are noisy: they always include observation noise and, depending on robustness trials, can be biased. Data are used as inputs in control rules for managing removals.

## Results

A Shiny application for visualizing results is available at `https://pelabox.univ-lr.fr/pelagis/DART/`. Results will be detailed and discussed during the talk. The main one is shown on Figure 2.

## Conclusion

Future research directions will close the talk. This research has been published in *PeerJ* (Ouzoulias et al. 2024).

# Bibliographie

Bordet, C. and Rivest, L.-P. (2014), A Stochastic Pella Tomlinson Model and Its Maximum Sustainable Yield, *Journal of Theoretical Biology*, 360, pp. 46-53

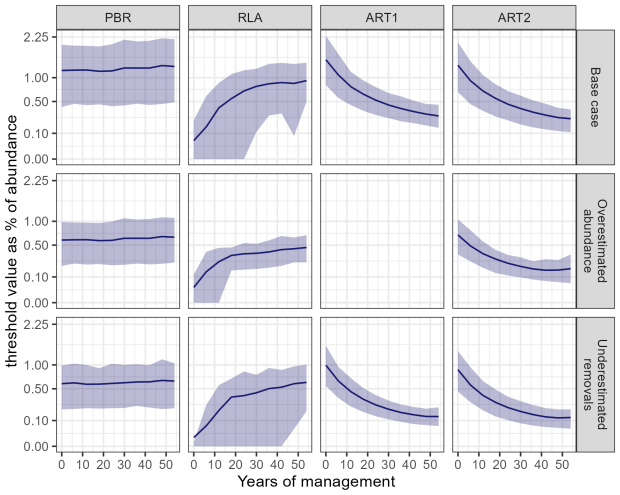



Figure 2: **Different paths to managing removals to ensure long-term population viability** The behaviour of the removals limit set by the different control rules was very different, with only ART achieving long-term viability and to minimize removals limits over time. Average and 80% confidence bands (from the 100 simulations) are depicted.


Bousquet, N. and Duchesne, T. and Rivest, L. P. (2008) Redefining the Maximum Sustainable Yield for the Schaefer Population Model Including Multiplicative Environmental Noise, *Journal of Theoretical Biology*, 254(1), pp. 65-75

Cook, J. and Fúquene, J. and Pericchi, L. (2011), Skeptical and Optimistic Robust Priors for Clinical Trials, *Revista Colombiana de Estadistica*, 34(2), pp. 333-345

Cooke, J. G. (1999) Improvement of Fishery-Management Advice Through Simulation Testing of Harvest Algorithms, *ICES Journal of Marine Science*, 56, pp. 797-810

Genu, M. and Gilles, A. and Hammond, P. and Macleod, K. and Paillé, J. and Paradinas, I. A. and Smout, S. and Winship, A. and Authier, M. (2021) Evaluating Strategies for Managing Anthropogenic Mortality on Marine Mammals: an R Implementation with the Package `RLA`, *Frontiers in Marine Science*, 8, pp. 795953

Ouzoulias, F. and Bousquet, N. and Genu, M. and Gilles, A. and Spitz, S. and Authier, M. (2024), Development of a New Control Rule for Managing Anthropogenic Removals of Protected, Endangered or Threatened Species in Marine Ecosystems, *PeerJ*, 12, pp. e16688

Pella, J. J. and Tomlinson, P. K. (1969), A Generalized Stock Production Model, *Inter-American Tropical Tuna Commission Bulletin*, 13(3) pp. 416-497

Rayner, S. (2012) Uncomfortable Knowledge: the Social Construction of Ignorance in Science and Environmental Policy Discourses, *Economy and Society*, 41(1), pp. 107-125

Schweder, T. (2000) Distortion of Uncertainty in Science: Antarctic Fin Whales in the 1950s, *Journal of International Wildlife Law and Policy*, 3(1), pp. 73-92

Wade, P. R. (1998), Calculating Limits To the Total Allowable Human-Caused Mortality of Cetaceans and Pinnipeds, *Marine Mammal Science*, 14(1), pp. 1-37

# Estimation consistante du nombre de clusters non vides dans les modèles de mélange bayésiens par régression sur profils d'exposition. Application en épidémiologie des rayonnements ionisants

Julie Fendler [1] & Sophie Ancelet [1] & Chantal Guihenneuc [2]

[1] *IRSN, PSE-SANTE/SESANE/LEPID, France,*
*emails: julie.fendler@irsn.fr ; sophie.ancelet@irsn.fr*
[2] *Université de Paris Cité, BioSTM - UR 7537, Paris, France,*
*email: chantal.guihenneuc-jouyaux@u-paris.fr*

**Résumé.** Depuis sa définition par Wild (2005), l'exposome est de plus en plus mis en avant pour son rôle dans l'apparition et le développement de maladies multifactorielles telles que le diabète, le cancer ou l'asthme. Cependant, les données d'exposome sont des données de grande dimension, souvent composées de variables fortement corrélées entre elles (multi-colinéarité). Pour l'analyse de données multi-colinéaires, des méthodes statistiques spécifiques sont nécessaires. Parmi elles, les modèles de mélange bayésiens par régression sur profils d'exposition (BPRM). Ces modèles hiérarchiques permettent de former des groupes d'individus ayant des profils d'exposition similaires à plusieurs facteurs de risque et d'estimer un risque sanitaire pour chacun de ces groupes. La répartition des individus en groupes et l'estimation du risque sanitaire associé se font conjointement sous le paradigme bayésien. Dans les modèles BPRM, le sous-modèle permettant l'attribution de chaque individu à un groupe repose sur un processus de Dirichlet latent, pour lequel l'estimation du nombre de groupes non vides est connue pour être inconsistante. Notre cas d'application demandant d'estimer un nombre interprétable de groupes non vides en présence de données faiblement informatives, nous proposons d'adapter l'algorithme MCMC d'inférence desmodèles BPRM afin de faciliter l'estimation du paramètre de concentration du processus de Dirichlet latent, dans une situation où le signal dans les données est faible. Enfin, nous adaptons aux modèles BPRM et nous comparons différentes méthodes de post-traitement afin de permettre la consistance du nombre estimé de groupes non vides.

**Mots-clés.** Modèles de mélange, Multi-colinéarité, Processus de Dirichlet, Inférence bayésienne, Analyse de survie, Epidémiologie

**Abstract.** Since its definition by Wild (2005), the exposome has been increasingly highlighted for its role in the onset and development of multifactorial diseases such as diabetes, cancer and asthma. However, exposome data are large-scale data composed of variables that are often highly correlated with each other (multi-colinearity). Specific statistical methods are needed to deal with multi-colinear data. These include Bayesian Profiles Regression Mixture models (BPRM). These hierarchical models allow to identify clusters of individuals with similar exposure profiles to several risk factors and to estimate a

health risk associated to each cluster. Clustering and risk estimation are carried out jointly under the Bayesian paradigm. In BPRM models, the sub-model used to assign each individual to a cluster is based on a latent Dirichlet process, that is well-known to provide an inconsistent estimation of the number of non-empty clusters. Since our application requires the estimation of an interpretable number of non-empty clusters in the presence of poorly informative data, we propose to adapt the MCMC algorithm used to fit the BPRM models in order to make easier the estimation of the concentration parameter of the latent Dirichlet process, when the data signal is weak. Finally, we adapt to BPRM models and we compare various post-processing methods in order to allow consistency in the estimated number of non-empty groups.

**Keywords.** Mixture models, Multi-colinearity, Dirichlet process, Bayesian Inference, Survival Modeling, Epidemiology

# 1 Introduction

Nous nous intéressons aux méthodes statistiques permettant de traiter le problème dit de multicolinéarité, posé par la prise en compte simultanée de variables explicatives fortement corrélées entre elles dans des modèles de régression multiples. Ce problème se rencontre notamment dans les études épidémiologiques, dont l'un des enjeux actuel est de tenir compte au mieux de l'exposome, défini par Wild (2005), lorsqu'on estime des risques de maladies multifactorielles telles que le diabète, le cancer ou l'asthme.

Dans le cas d'une multicolinéarité trop prononcée, l'utilisation d'un modèle de régression linéaire standard incluant simultanément plusieurs facteurs de risque (i.e., variables explicatives) corrélés n'est pas adaptée: les coefficients de régression estimés sont instables et peuvent changer de signe en fonction du sous-ensemble de variables explicatives sélectionné mais aussi du sous-échantillon d'individus statistiques considéré. L'imprécision des estimations se traduit par une variance augmentée ne permettant souvent plus de conclure à des associations significatives entre la variable réponse et les variables explicatives. Dans la pratique, ce problème est souvent surmonté en sélectionnant uniquement une variable explicative (celle influençant le plus la variable réponse) parmi les variables corrélées disponibles. Cette solution ne permet ni l'étude des effets de l'ensemble des variables explicatives d'intérêt ni l'étude de probables effets sanitaires conjoints (i.e., synergiques ou antagonistes).

Les modèles de mélange bayésiens par régression sur profils d'exposition (modèles BPRM) proposés par Molitor *et al.* (2010) permettent de répondre au problème de multicolinéarité ci-dessus. Ces modèles hiérarchiques permettent de former des groupes d'individus ayant des profils d'exposition similaires à plusieurs facteurs de risque et d'estimer un risque sanitaire pour chacun de ces groupes. La

répartition des individus en groupes et l'estimation du risque sanitaire associé se font conjointement sous le paradigme bayésien. Un post-traitement (i.e., mené après l'inférence) est indispensable à l'identification et la caractérisation des clusters.

Dans les modèles BPRM, le sous-modèle permettant l'attribution de chaque individu à un groupe repose sur un processus de Dirichlet latent, pour lequel l'estimation du nombre de groupes non vides est connue pour être inconsistante (Miller et Harrisson (2013)). Par ailleurs, l'inférence bayésienne des modèles BPRM, et notamment du paramètre de concentration du processus de Dirichlet latent, à l'aide d'un algorithme MCMC standard, pose des difficultés (Liverani *et al.* (2015); Belloni *et al.* (2020)).

Dans ce travail, le cas d'application traité (cf. section 2) présente deux spécificités : 1) il demande d'estimer un nombre interprétable de clusters non vides; 2) il demande d'estimer un risque sanitaire faible en présence de données potentiellement peu informatives. Dans ce contexte, nous proposons tout d'abord d'adapter l'algorithme d'inférence bayésienne proposé par Liverani *et al.* (2015) afin de faciliter l'estimation du paramètre de concentration du processus de Dirichlet latent, dans une situation où le signal dans les données est potentiellement faible. De plus, nous proposons d'adapter les méthodes de post-traitement proposées par Wade et Ghahramani (2018) et Guha *et al.* (2019) pour les processus de Dirichlet aux modèles BPRM, afin de permettre la consistance du nombre estimé de clusters non vides. Enfin, nous comparons ces deux méthodes à celles proposées par Molitor *et al.* (2010) et Belloni *et al.* (2020). A notre connaissance, aucune étude n'a été menée afin de comparer ces différentes méthodes.

## 2 Cas d'étude

La population des mineurs d'uranium est une population de référence pour étudier les effets sanitaires d'une exposition chronique à faibles doses et à différentes sources de rayonnements ionisants (RI). En effet, dans le cadre de leur activité professionnelle, les mineurs d'uranium sont simultanément exposés au radon et ses descendants à vie courte (appelé simplement radon par la suite), aux poussières d'uranium et aux rayonnements gamma. En pratique, les effets sanitaires associés à ces expositions sont généralement étudiés de manière monofactorielle pour chaque source de RI. Ces estimations servent ensuite de base à la définition de normes de radio-protection.

Dans ce travail, nous cherchons à estimer et caractériser le risque de décès par cancer du poumon dans la cohorte française post-55 des mineurs d'uranium, en tenant compte de leur exposition simultanée au radon, aux rayonnements gamma et aux poussières d'uranium. Cette cohorte inclut 3377 mineurs d'uranium embauchés après le 31 décembre 1955. Ils ont été suivis en moyenne pendant 37 ans et 130 cas de décès par cancer du poumon ont été observés. De précédents

travaux ont montré que ces expositions sont fortement corrélées entre elles et qu'on est bien dans un contexte de multicolinéarité prononcée (Belloni *et al.* (2020); Vacquier et al. (2011)). Chacune de ces trois sources d'exposition aux RI a été associée, de manière individuelle, à une augmentation du risque de décès par cancer du poumon (Rage *et al.* (2015)). Cependant l'impact d'une co-exposition simultanée à l'ensemble de ces sources est encore mal caractérisé.

# 3 Méthode

## 3.1 Description du modèle BPRM

Un modèle de mélange bayésien par régression sur profils d'exposition (BPRM) peut être vu comme un modèle hiérarchique composé de trois sous-modèles :

- un sous-modèle de maladie qui décrit l'association entre la survenue d'une pathologie (exprimée, par exemple, par l'âge au décès par cancer du poumon d'un mineur) et un profil d'exposition. La force de cette association est quantifiée avec un coefficient de risque associé à chaque profil et un risque instantané de base de développer la pathologie considérée.
- un sous-modèle d'exposition qui décrit la distribution de probabilité des variables d'exposition (continues et discrètes) dans chaque groupe
- un sous-modèle d'attribution qui décrit la répartition des individus dans les différents groupes

### 3.1.1 Sous-modèle de maladie

Dans ce travail, le sous-modèle de maladie considéré est un modèle de survie. Nous adaptons aux modèles BPRM, le modèle exposition-risque linéaire - appelé modèle en excès de risque instantané (EHR) - classiquement utilisé en épidémiologie des RI.

La variable réponse $E_i$ est l'âge (en jours) à la survenue de l'événement considéré d'un individu $i$. Cette variable est censurée à droite (individu perdu de vue, décès par autre cause, fin de suivi) et tronquée à gauche (car l'échelle de temps considérée est l'âge). On note $Z_i$ l'âge de l'individu $i$ à la censure. On observe $Y_i = min(E_i, Z_i)$ et $\delta_i^Y$ l'indicateur binaire de non-censure (=1 si l'événement étudié est observé pour l'individu $i$; =0 si l'individu $i$ est censuré avant la survenue de l'événement étudié).

Le risque instantané de l'individu $i$ au temps $t$, noté $h_i(t)$, est défini par:

$$h_i(t) = h_0(t) \cdot (1 + \beta_{C_i})$$

$C_i$ désigne le label de groupe (inconnu) auquel appartient l'individu $i$ et $\beta_c$ l'excès de risque instantané (inconnu) de survenue de l'événement étudié, associé au groupe $c$. $h_0(t)$ désigne le risque instantané de base de survenue de l'événement considéré au temps $t$. Nous supposons une fonction de risque instantanée de base $h_0$ de type Weibull :

$$h_0 : t \mapsto \xi t^{\nu-1}$$

définie par deux paramètres inconnus: un paramètre d'échelle $\xi > 0$ et un paramètre de forme $\nu > 1$ permettant d'assurer que le risque instantané de base soit positif et croissant avec le temps.

### 3.1.2 Sous-modèle d'exposition

Le sous-modèle d'exposition décrit la distribution de probabilité des variables d'exposition (continues et discrètes) dans un groupe $c \in \mathbb{N}^*$.

Pour notre cas d'étude, les différentes variables d'exposition considérées pour la caractérisation des groupes et profils d'exposition des mineurs d'uranium français sont les suivantes :

- l'exposition au radon du mineur $i$, $X_i^R$, cumulée au cours de sa carrière (variable continue);
- l'exposition aux rayonnements gamma du mineur $i$, $X_i^G$, cumulée au cours de sa carrière (variable continue);
- l'exposition aux poussières d'uranium du mineur $i$, $X_i^P$, cumulée au cours de sa carrière (variable continue);
- le poste de travail $J_i$ du mineur $i$. Cette variable est un proxy pour les conditions d'exposition ainsi que d'éventuelles autres expositions professionnelles (variable catégorielle à 5 modalités);
- l'âge à la première exposition $A_i$ du mineur $i$ (variable continue);
- la localisation de la mine $M_i$ du mineur $i$. Cette variable est une proxy des conditions de travail du mineur car en fonction du type de sol, les techniques d'exploitation du minerai diffèrent (variable catégorielle à deux modalités);
- la durée d'exposition $T_i$ du mineur $i$ (variable continue).

L'ensemble $(X_i^R, X_i^G, X_i^P, A_i, J_i, M_i, T_i)$ constitue le profil d'exposition du mineur $i$. Chaque variable d'exposition suit une loi de probabilité conditionnelle au groupe $c$. On note $X_i^{cat,k}$ la $k^{\text{i-ème}}$ variable catégorielle du profil d'exposition du mineur $i$ et $X_i^{cont,k}$ la $k^{\text{i-ème}}$ variable continue du profil d'exposition du mineur $i$. On suppose que :

- $X_i^{cat,k}|C_i = c \sim \mathcal{M}ultinomial(\boldsymbol{p}_c^{cat,k})$;
- $X_i^{cont,k}|C_i = c \sim \mathcal{LN}(\mu_c^{cont,k}, \sigma_c^{cont,k})$

avec $\mathcal{LN}$ la loi lognormale.

### 3.1.3 Sous-modèle d'attribution

Le sous-modèle d'attribution répartit les individus dans les différents groupes. Il est défini par:

$$P(C_i = c) = \phi_c \qquad \text{pour tout} \qquad \text{c} \in \mathbb{N}^*$$

Le vecteur $\boldsymbol{\phi} = (\phi_c)_{c\in\mathbb{N}^*}$ definit la probabilité d'appartenance des individus à chacun des groupes $c$. Ce vecteur suit un processus de Dirichlet. La construction de ces poids de mélange, appelée "stick-breaking", est définie grâce aux relations suivantes:

$$\phi_c = V_c \cdot (1 - \sum_{k=1}^{c-1} \phi_k) \qquad \text{pour tout} \qquad \text{c} \in \mathbb{N}^*$$
$$\phi_1 = V_1$$

avec $(V_c)_{c\in\mathbb{N}^*}$ des variables aléatoires latentes inépendantes définies, pour tout c, par:

$$V_c \sim Beta(1, \alpha)$$

Si, en théorie, il existe une infinité de groupes d'individus différents, en pratique, il existe un nombre fini de groupes non vides $C$. Le vecteur $\boldsymbol{\phi}$ est donc un vecteur sparse avec $\phi_c = 0$ pour tout $c > C$. La valeur de $C$ dépend du paramètre dit de concentration $\alpha$. Plus la valeur de $\alpha$ est petite, plus le nombre estimé de groupes non vides est petit et vice versa.

## 3.2 Choix des lois *a priori*

Le Graphe Acyclique Orienté de notre modèle est donné dans la Figure 1.

Les lois *a priori* considérées sont les suivantes :

- $\beta_c \sim \mathcal{N}(0, 10^6)$, $c \in \mathbb{N}^*$;
- $\mu_c^{cont,k} \sim \mathcal{N}\left(\mu_{prior}^{(k)}, \sigma_{prior}^{2(k)}\right)$, $c \in \mathbb{N}^*$. Pour l'âge à la première exposition et la durée d'exposition, les lois *a priori* normales considérées sont centrées et de grande variance $10^6$. Pour les expositions au radon, aux rayonnements gamma et aux poussières d'uranium, les lois *a priori* sont des lois normales dont les valeurs des paramètres ont été définies à partir de la distribution de ces expositions dans la cohorte allemande des mineurs d'uranium;

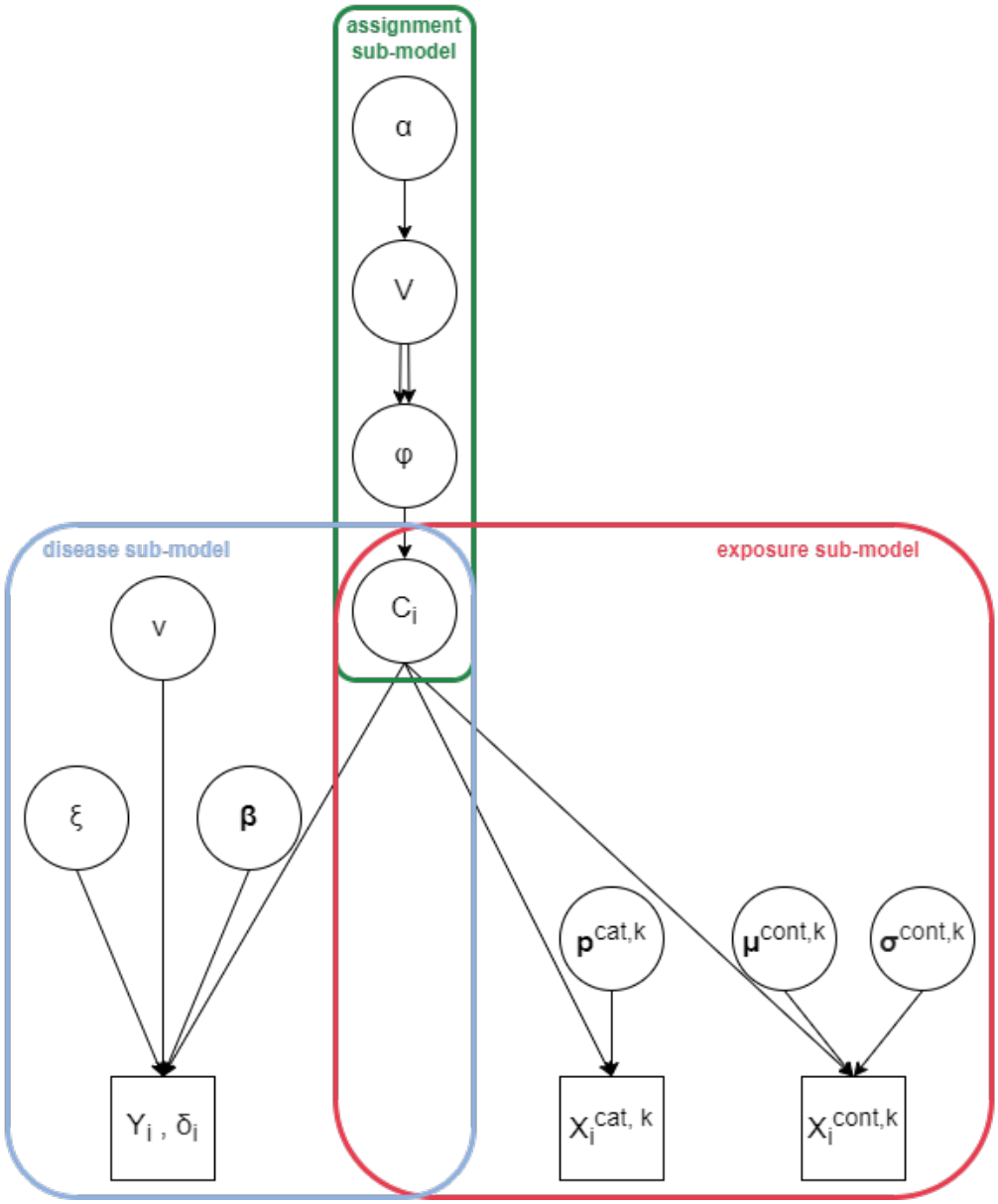



Figure 1: Graphe Acyclique Orienté du modèle de mélange bayésien par régression sur profils d'exposition.

- $\nu' = \nu - 1 \sim \mathcal{G}amma(0.001, 0.001)$;
- $\xi' = \xi \times 10^{25} \sim \mathcal{G}amma(1, 1)$;
- $\sigma_c^{cont,k} \sim \mathcal{G}amma(0.001, 0.001)$, $c \in \mathbb{N}^*$;
- $\boldsymbol{p}_c^{cat,k} \sim \mathcal{D}irichlet(0.5, 0.5, \ldots)$, $c \in \mathbb{N}^*$.
- $\alpha \sim \mathcal{G}amma(1, 2)$ tel que recommandé dans Liverani *et al.* (2015).

### 3.3 Inférence bayésienne

Un algorithme Monte-Carlo par Chaînes de Markov (MCMC) de type Metropolis-within Gibbs adaptatif a été implémenté en Python 3.11. Par ailleurs, le nombre de variables de groupe latentes étant infini, un slice sampler a été utilisé afin de se ramener à un nombre de variable latente fini à chaque itération, tel que proposé par Walker (2007). Enfin, la loi de probabilité *a posteriori* jointe étant potentiellement multimodale, ce qui pourra notamment être favorisé en présence de données peu informatives, l'estimation du paramètre de concentration $\alpha$ du processus de Dirichlet latent pose des difficultés (Belloni *et al.* (2020)). Afin de faciliter la convergence l'algorithme, nous avons couplé notre algorithme MCMC avec un algorithme de type "parallel tempering".

### 3.4 Post-traitement

Un post-traitement est indispensable à l'identification et la caractérisation des clusters. En effet, à chaque itération de l'algorithme, une partition différente de l'ensemble des individus en groupes est estimée. Le post-traitement permet d'agréger ces différentes partitions.

Plusieurs méthodes de post-traitement sont comparées :

- La première est la méthode *ad hoc* proposée par Molitor *et al.* (2010). Elle consiste à définir les groupes grâce à un algorithme K-meloids prenant comme paramètre d'entrée la matrice de dissimilarité moyenne (i.e. moyenne des matrices de dissimilarité obtenues à chaque itération).
- La deuxième méthode est proposée par Wade et Ghahramani (2018). La partition optimale est la partition minimisant, sur l'ensemble des partitions possibles, une fonction de perte appelée "variation of information".
- La troisième méthode agrège l'algorithme Merge-Truncate-Merge de Guha *et al.* (2019) et la méthode de post-traitement proposée par Belloni *et al.* (2020). Le premier algorithme permet de regrouper à chaque itération les groupes de petites tailles créés artificiellement par le processus de Dirichlet. Cette étape permet de garantir la consistance du processus de Dirichlet quant au nombre estimé de groupes non-vides (Guha *et al.* (2019))).

La seconde étape consiste à trouver la partition dont la matrice de dissimiliarité minimise la distance des moindres carrés à la matrice de dissimilarité moyenne.

Les lois *a posteriori* des paramètres définissant chaque groupe i.e., $\theta_c = (\boldsymbol{\mu}_c^{cont}, \boldsymbol{\sigma}_c^{cont}, \boldsymbol{p}_c^{cat}))$ sont alors déduites de cette meilleure partition $\boldsymbol{z}^{best}$. Ainsi, un échantillon *a posteriori* du paramètre $\theta_c$ du groupe $c$ est obtenu à chaque itération $j$ par $\bar{\theta}_{c,j} = \frac{1}{n_c} \sum_{i:z_i^{best}=c} \theta_{z_i^j,j}$ avec $n_c$ le nombre d'individus du groupe $c$, $z_i^j$ le groupe auquel appartient l'individu $i$ à l'itération $j$.

# 4 Résultats

Une étude par simulations sera présentée afin de déterminer, 1) l'impact de l'algorithme "parallel tempering", 2) l'impact de la procédure de post-processing, sur le nombre estimé de groupes non vides des modèles BPRM.
Une application du modèle proposé aux données de la cohorte post-55 des mineurs d'uranium français sera également présentée.

# Bibliographie

Wild C. P. (2005), Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology, Cancer Epidemiol Biomarkers, 14(8), 1847-50
Miller J. W. and Harrison M. T. (2013), Inconstistency of Pitman-Yor process mixtures for the number of components, ArXiv, 1309.0024v1
Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., & Richardson, S. (2015). PReMiuM: An R package for profile regression mixture models using Dirichlet processes. Journal of statistical software, 64(7), 1.
Molitor, J., Papathomas, M., Jerrett, M., & Richardson, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics, 11(3), 484-498.
Rage, E., Caër-Lorho, S., Drubay, D., Ancelet, S., Laroche, P., & Laurier, D. (2015). Mortality analyses in the updated French cohort of uranium miners (1946–2007). International archives of occupational and environmental health, 88(6), 717-730.
Vacquier, B., Rage, E., Leuraud, K., Caër-Lorho, S., Houot, J., Acker, A., & Laurier, D. (2011). The influence of multiple types of occupational exposure to radon, gamma rays and long-lived radionuclides on mortality risk in the French "post-55" sub-cohort of uranium miners: 1956–1999. Radiation research, 176(6), 796-806
Wade S., & Gahramani Z., (2018), Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion), Bayesian Analysis, 13(2), 556-629
Guha A., Ho N., & Nguyen X. (2019) On posterior contraction of parameters and interpretability in Bayesian mixture modeling, arXiv, 1901.05078v1

Belloni M., Laurent O., Guihenneuc C., & Ancelet S. (2020) Bayesian Profile Regression to Deal With Multiple Highly Correlated Exposures and a Censored Survival Outcome. First Application in Ionizing Radiation Epidemiology, frontiers in Public Health, 8, 557006

# Inférence causale

# Découverte de causalité pour séries temporelles en présence de causes cachées

Antonin Arsac [†,1] & Aurore Lomet [‡,2] & Jean-Philippe Poli [†,3]

[†] *Université Paris-Saclay, CEA LIST, F-91120, Palaiseau, France*
[‡] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191, Gif-sur-Yvette, France.*
[1] *antonin.arsac@cea.fr,* [2] *aurore.lomet@cea.fr,* [3] *jean-philippe.poli@cea.fr*

**Résumé.** La recherche de causalité vise à définir les relations de causes à effets entre différentes variables à partir de données. Cet apprentissage peut s'avérer difficile notamment lorsque les observations d'un système complexe sont incomplètes et que l'intégralité des causes d'un phénomène ne sont pas observées. A cette complexité, le cadre spécifique des séries temporelles ajoute la nécessité de considérer des structures de dépendance telles que l'auto-corrélation. Le travail présenté dans ce papier se concentre sur l'apprentissage d'un graphe causal à partir de données de séries temporelles issues d'un système complexe en présence de causes latentes. Dans ce cadre, nous proposons une méthode permettant d'identifier des liens entre les séries, qu'ils soient linéaires ou non-linéaires et possiblement décalés dans le temps, pour construire un graphe causal. Pour ce faire, nous combinons un algorithme de découverte causale et une mesure basée sur la théorie de l'information. L'approche est évaluée sur un jeu de données simulées issues de la littérature, démontrant sa pertinence dans une étude comparative.
**Mots-clés.** Découverte causale, séries temporelles, théorie de l'information, variables latentes

**Abstract.** Causality aims to define cause-and-effect relationships between different variables based on data. Learning causality can be challenging, particularly when observations of a complex system are incomplete and all the causes of a phenomenon are not observed. To this complexity, the specific framework of time series data adds the necessity to consider dependency structures such as auto-correlation. The work presented in this paper focuses on learning a causal graph from time series data generated by a complex system in the presence of latent causes. In this context, we propose a method to identify relationships between the series, whether linear or non-linear and possibly time-lagged, to construct a causal graph. To achieve this, we combine a causal discovery algorithm with an information-theoretic measure. The approach is evaluated on simulated data from the literature, demonstrating its relevance in a comparative study.
**Keywords.** Causal discovery, time series, Information theory, latent variables

## 1 Introduction

L'étude de systèmes complexes évoluant dans le temps peut requérir l'analyse de séries temporelles. Ces données sont présentes dans de nombreux domaines comme la médecine

(au travers d'EEG par exemple), les analyses de marché [Brodersen et al., 2015] ou encore dans l'étude du climat [Runge et al., 2019a]. Explorer les relations de causes à effets entre différentes variables temporelles permet de comprendre au mieux ces phénomènes. En effet, simplement observer des corrélations décalées dans le temps peut s'avérer insuffisant : ces dernières peuvent provenir de variables cachées (variables latentes) et peuvent être très faibles dans le cas de systèmes non-linéaires bruités. De récents développements dans le domaine de la causalité, comme ceux de [Pearl, 2009], permettent la distinction entre fausse corrélation et relation causale, au travers des graphes dirigés acycliques (DAG). Ainsi, l'objectif de ce travail est de construire un graphe causal à partir de données de séries temporelles multivariées en présence de variables non-observées.
Dans les graphes causaux, chaque noeud correspond à une variable et la présence d'une arête orientée entre deux noeuds représente la relation d'une cause vers son effet. Une absence d'arête entre deux noeuds se traduit par une indépendance ou une indépendance conditionnelle dans les données [Spirtes et al., 2000]. Cependant, un grand nombre de modèles et de méthodes reposent sur des hypothèses souvent peu réalistes comme la linéarité, une distribution *a priori* des données ou encore de suffisance causale, indiquant que toutes les causes de chaque effet sont observées.
Pour répondre à ces problèmes, les travaux présentés dans ce papier visent à inférer les relations causales entre des séries temporelles avec peu d'hypothèses. Pour ce faire, nous combinons un algorithme de découverte de causalité, l'algorithme FCI [Spirtes et al., 2000, Zhang, 2008] avec une mesure d'indépendance conditionnelle, la *Partial Mutual Information from Mixed Embedding* (PMIME) [Kugiumtzis, 2013]. L'algorithme FCI, couplé à un test d'indépendance conditionnelle, permet de construire un graphe causal en prenant en compte l'existence de variables non-observées. La mesure non-paramétrique PMIME, basée sur la théorie de l'information, développée pour les séries temporelles, considère l'auto-corrélation au sein des variables.

Dans cet article, nous introduisons en partie II les notions nécessaires pour traiter la causalité au sein des séries temporelles. La partie III présente la méthode que nous proposons. Enfin, nous réalisons des expérimentations dans la partie IV avant d'en discuter les résultats et de conclure en partie V.

# 2 Découverte causale et séries temporelles

## 2.1 Modèles de graphe

La causalité entre séries temporelles peut être représentée par des réseaux causaux bayésiens, une classe de modèles de graphe permettant une représentation probabiliste de variables aléatoires.
Le principe de priorité temporelle, qui stipule qu'une cause précède ses effets, induit l'asymétrie de la causalité dans le temps. Elle permet ainsi d'orienter les relations causales dans un graphe lorsqu'une cause est déjà connue. Si une variable prise à un temps $t - \tau$ cause une autre variable à un temps $t$, la relation causale est définie comme une relation causale décalée. Si la cause apparaît en même temps que son effet, la relation est dite instantanée.

Il existe plusieurs types de graphes pour représenter les relations causales entre séries temporelles comme le graphe causal fenêtre (*window causal graph*), Figure 1 (a) et le graphe causal résumé (*summary causal graph*), Figure 1 (b). Dans le premier type de graphe, un noeud représente une variable à un instant donné, tandis que le second ne donne que la variable : l'information temporelle est donc perdue. De telles représentations sont possibles grâce à la stationnarité causale [Runge, 2018] qui stipule que toutes les relations causales restent constantes en direction au cours du temps.
Un noeud $X^r$ tel que $X^i \to X^r \leftarrow X^j$ est appelé un *collider* comme par exemple le noeud

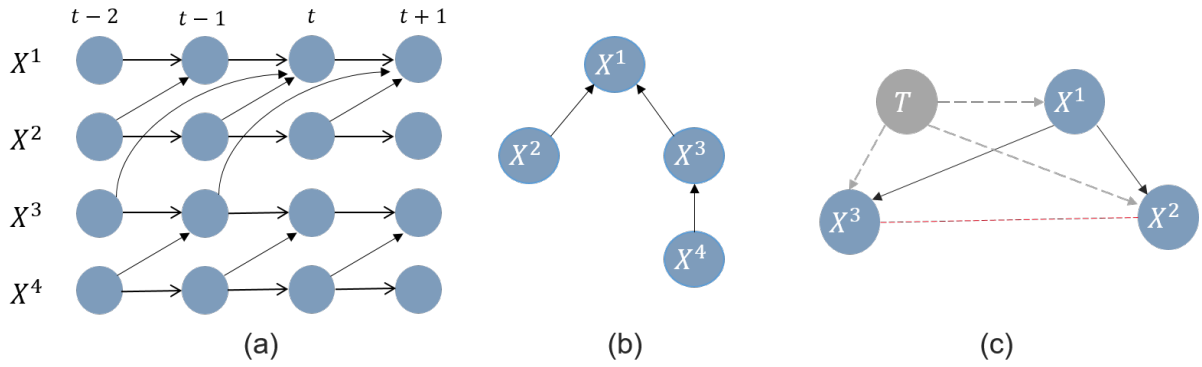



FIGURE 1 – Sur la gauche (a), un *window causal graph* et au milieu (b), le graphe causal *summary* correspondant. A droite (c), un autre *summary* où $T$ est un facteur confondant non-observé.

$X^1$ sur la Figure 1 (b). Les antécédents d'un noeud dans le graphe sont appelés ses parents, $X^1$ a pour parents le couple $(X^2, X^3)$. Un noeud $X^r$ appartient à l'ensemble de séparation de deux noeuds $(X^i, X^j)$ si conditionner sur le premier rend le couple indépendant, par exemple le noeud $X^3$ d-sépare $X^4$ et $X^1$ dans la Figure 1 (b). Si une variable cause plusieurs noeuds, elle est appelée une cause commune (*confounder*), comme le noeud $X^1$ sur la Figure 1 (c). Une cause commune cachée est un *hidden confounder*, c'est le cas de la variable $T$ sur cette même figure.
Pour constituer les arêtes du graphe, les critères d'indépendance ou d'indépendance conditionnelle sont primordiaux. Ils font en effet le lien entre des paramètres statistiques évalués dans les données et le graphe causal, par la propriété causale de Markov et la propriété de *faithfulness* [Peters et al., 2017]. Retrouver ces liens causaux est appelé la recherche de causalité (*causal discovery*). Il existe un grand nombre de méthodes de découverte causale catégorisées en différentes familles selon leur approche pour construire le graphe [Vowels et al., 2022] dont celle dite *constraint-based*. Cette dernière cherche à établir un graphe qui respecte au mieux les ensembles d'indépendance conditionnelle détectés dans les données.

## 2.2 Le problème des variables latentes

Dans la plupart des situations réelles, il est difficile d'affirmer qu'un jeu de données contient toutes les causes du système observé. Cela implique donc l'existence de variables latentes, ou non-observées. Ignorer ces variables peut conduire à des conclusions fausses et

des analyses incomplètes car elles peuvent introduire un biais significatif dans les résultats. Intégrer ces variables cachées dans les modèles permet donc de mieux étudier les mécanismes sous-jacents du système.
Pour cela, les DAG sont étendus pour contenir jusqu'à six types d'arêtes : une arête non-dirigées ($-$), une flèche seule ($\rightarrow$), une double flèche ($\leftrightarrow$), une arête non-dirigée d'un côté et indéterminée de l'autre ($\circ\!\!\rightarrow$) et enfin une arête indéterminée des deux côtés ($\circ\!\!-\!\!\circ$). Le graphe ainsi inféré est appelé un *Partial Ancestral Graph* (PAG). Dans un PAG, $X^1 \rightarrow X^2$ signifie que $X^1$ est une cause (potentiellement indirecte) de $X^2$ et que $X^2$ ne cause pas $X^1$. Ainsi, une variable en cause une autre si elle est un antécédent de cette variable. Une double flèche entre deux noeuds indique qu'il n'y a pas de relation causale directe entre eux mais qu'une cause commune les influence. Munis de ces relations, certains algorithmes ont été établis sans l'hypothèse de suffisance causale. Parmi ceux-ci, FCI [Spirtes et al., 2000, Zhang, 2008], un algorithme *constraint-based*, utilise les critères d'indépendance (conditionnelle) pour estimer un graphe causal et un ensemble de règles pour orienter ce graphe et détecter des variables latentes.
Une règle fondamentale, nommée *origin of causality*, permet d'orienter les triplets de variables $X^i \!*\!\!-\!\!\circ X^r \circ\!\!-\!\!* X^j$ tels que $X^i$ et $X^j$ ne sont pas adjacents, en *unshielded colliders*. Un *unshielded collider* est un noeud $X^r$ tel que $X^i \!*\!\!\rightarrow X^r \!\leftarrow\!\! * X^j$, si $X^r$ n'est pas dans l'ensemble de séparation de $X^i$ et $X^j$. Cette règle, souvent mentionnée comme la règle $R_0$, permet de former des ensembles de *Possible D-separation* [Spirtes et al., 2000] entre deux noeuds. Ces ensembles, présents dans les PAG, se comportent de manière similaire aux parents de noeuds dans un DAG.

## 2.3 Séries temporelles

Les séries temporelles sont des données ordonnées qui présentent souvent des structures d'auto-corrélation. Cela signifie que leur passé influence leur présent et leur futur. Il est nécessaire de les prendre en compte lors de la construction du graphe causal. Le décalage temporel entre deux observations permet d'orienter le graphe : une cause ne peut intervenir avant son effet. Le temps peut être vu comme un facteur confondant entre plusieurs variables : si chaque variable entretient une relation linéaire avec le temps alors elles partagent un point commun. Cela peut donc créer un lien entre ces variables, comme observé sur la Figure 1 (c). Sur cette dernière, le temps est représenté par la variable $T$ et agit sur les séries $X^1, X^2$ et $X^3$. Si $T$ n'est pas considérée dans le modèle, cela peut induire une corrélation entre ces variables et créer un lien entre $X^2$ et $X^3$ qui ne devrait pas exister.
Plusieurs algorithmes ont été proposés pour construire un graphe causal à partir de séries temporelles comme VarLiNGAM [Hyvärinen et al., 2010], PCMCI [Runge et al., 2019b], PCGCE [Assaad et al., 2022a] ou encore PC-PMIME [Arsac et al., 2023]. Plusieurs de ces méthodes s'appuient sur l'indépendance conditionnelle.

## 2.4 Tester l'indépendance conditionnelle

Deux variables aléatoires $X$ et $Y$ sont conditionnellement indépendantes à une troisième variable $Z$ si et seulement si (ssi) $P_{X,Y|Z} = P_{X|Z}P_{Y|Z}$, on note $X \perp\!\!\!\perp Y \mid Z$. Intuitivement, cela signifie que la distribution de $X$ sachant $Y$ et $Z$ est en fait complètement déterminée par la valeur de $Z$ seule, $Y$ devient superflue lorsque $Z$ est donnée.
Plusieurs tests d'indépendance conditionnelle sont basés sur la théorie de l'information qui permet d'établir des tests non-paramétriques et donc d'être appliquée à de nombreux modèles de données. Pour les séries chronologiques, des mesures ont été développées à partir de l'Information Mutuelle Conditionnelle (CMI) comme l'entropie de transfert partielle (PTE) [Vakorin et al., 2009] ou la mesure PMIME [Kugiumtzis, 2013].
La mesure PMIME est asymétrique, non-paramétrique et conçue pour détecter des couplages directs au sein de séries temporelles. Elle est dérivée d'un processus d'*embedding* basé sur un critère de sélection, la CMI. Dans le cas multivarié, pour évaluer si une variable $X$ a une influence sur une variable $Y$, conditionnellement à un ensemble de variables $\mathbf{Z} = (Z^1, ..., Z^{g-2})$, le processus construit itérativement un vecteur d'*embedding* $\mathbf{w}$ à partir des composantes décalées extraites de $(X, Y, \mathbf{Z})$ qui expliquent le mieux le futur de $Y$, noté $Y_t^T = (Y_{t+1}, ..., Y_{t+T})$. Plus précisément, le vecteur $\mathbf{w}$ est construit en piochant successivement des vecteurs dans la matrice $W = (X_{t-1}, ..., X_{t-\tau_{max}}, Y_{t-1}, ..., Y_{t-\tau_{max}}, Z^1_{t-1}, ..., Z^{g-2}_{t-\tau_{max}})$. Chaque itération définit un cycle d'*embedding* et utilise un critère d'arrêt, noté $A$, pour accepter ou refuser une composante. Une composante est acceptée si l'information qu'elle apporte améliore strictement l'information déjà contenue dans le vecteur d'*embedding*.
Ainsi, le vecteur $\mathbf{w}$ est formé de $g'$ variables décalées dans le temps sélectionnées par la CMI et peut se décomposer en $\mathbf{w}_t = (w_t^x, w_t^y, w_t^{\mathbf{Z}})$, où $w_t^x$ sont les composantes de $X$ sélectionnées dans le procédé, $w_t^y$, celles de $Y$ et les composantes restantes sont regroupées sous $w_t^{\mathbf{Z}}$. On quantifie alors l'effet causal de $X$ vers $Y$ conditionnellement à $\mathbf{Z}$ par un ratio entre une CMI et une information mutuelle :

$$R_{X\to Y|\mathbf{Z}} = \frac{I(Y_t^T, \mathbf{w}_t^x | \mathbf{w}_t^y, \mathbf{w}_t^{\mathbf{Z}})}{I(Y_t^T; \mathbf{w}_t)}.$$

La description détaillée de la construction du vecteur $\mathbf{w}$ est donnée dans [Kugiumtzis, 2013]. Si $w_t^x$ est vide cela signifie que $X$ n'a aucune influence sur $Y$, ce qui se transcrit sur la mesure $R$ : si $w_t^x$ est vide, alors $R$ est nulle. De plus, la mesure est bornée entre 0 et 1, 0 signifie indépendance et 1 signifie que le futur de $Y$ est totalement déterminé par $X$.

# 3 Approche proposée

## 3.1 Motivations et hypothèses

Dans ce travail nous considérons un système modélisé par des séries temporelles caractérisé par des relations linéaires ou non-linéaires entre les variables, sans faire d'hypothèse spécifique sur leur distribution *a priori*. Nous supposons l'existence de causes cachées communes (*hidden confounders*), écartant ainsi l'hypothèse de suffisance causale. Le modèle des séries tempo-

relles considéré est général, décrivant une variable $X^i$ observée au temps $t$ selon :

$$X_t^i = f(X_{t-\gamma_{1,1}}^{p_1}, X_{t-\gamma_{1,2}}^{p_1}, ..., X_{t-\gamma_{1,j_1}}^{p_1}, X_{t-\gamma_{2,1}}^{p_2}, ..., X_{t-\gamma_{k,j_k}}^{p_k}, \varepsilon_t^i),$$

où $p_1, ..., p_k$ sont les indices des variables parentes de $X^i$ et $\gamma_{l,m}$ représente le décalage $m$ avec lequel la variable $X^{p_l}$ influence $X_t^i$, tel que $\gamma_{l,1} > \gamma_{l,2} > ... \geq 0$. Les $(\varepsilon_t)_i^n$, indépendants les uns des autres et indépendants des causes de $X^i$, représentent un bruit ajouté.
Pour construire un graphe causal à partir de séries temporelles dans ce cadre, différentes méthodes existent comme tsFCI [Entner and Hoyer, 2010], une extension de FCI capable de traiter des séries temporelles. Cependant, ses auteurs indiquent que tsFCI n'est pas fiable pour pratiquement toutes les tailles de séries temporelles. D'autres extensions existent également comme SVAR-GFCI [Malinsky and Spirtes, 2018], qui se concentre sur les modèles auto-régressifs structurels linéaires ou LPCMCI [Gerhardus and Runge, 2020]. Ce dernier modèle semble respecter nos conditions de travail mais souffre d'un temps de calcul conséquent lors de son utilisation avec une mesure de dépendance non-linéaire, la CMI, et offre une faible interprétabilité.
Nous proposons donc d'utiliser un algorithme de type FCI avec la mesure d'indépendance conditionnelle PMIME. La méthode, nommée FCI-PMIME, permet de construire un graphe causal à partir de données de séries temporelles en l'absence de suffisance causale. Cet algorithme traite des relations linéaires et non-linéaires entre les variables. De plus, [Papana et al., 2013] montre que la mesure PMIME détecte mieux que la PTE, les dépendances et indépendances conditionnelles dans le cas de systèmes non-linéaires, ce qui s'avère crucial lors de l'utilisation d'une méthode *constraint-based*. Un avantage certain de la mesure PMIME est qu'elle prend en compte directement l'aspect temporel des variables, en considérant une fenêtre temporelle lors du test, contrairement à la CMI par exemple. Enfin, la mesure est bornée entre 0 et 1, simplifiant son interprétabilité et évitant l'ajout d'un test de signification statistique supplémentaire.

## 3.2 L'algorithme FCI-PMIME

L'algorithme FCI-PMIME fonctionne de façon analogue aux méthodes constraint-based. Il essaye dans un premier temps de trouver les ensembles d'indépendance et d'indépendance conditionnelle dans les données puis applique des règles d'orientation spécifiques pour établir le graphe causal. Une caractéristique essentielle de FCI-PMIME est sa capacité à mesurer des interactions décalées dans le temps ainsi que des relations instantanées. La mesure PMIME est utilisée pour évaluer les dépendances du passé vers le futur, tandis que pour les relations instantanées, l'algorithme recourt à la MI ou la CMI [Runge, 2018].
L'algorithme commence par former un graphe complet $G = (V, E)$, où $V = (V_{t-}, V_t)$ représente les noeuds du passé et du présent et $E$, les arêtes. Une phase d'élagage du graphe débute dans laquelle une arête entre $X$ et $Y$ est enlevée si $R_{X \to Y} = 0$, où $R$ est la mesure PMIME ou la CMI. Lorsque toutes les arêtes ont été testées et que certaines ont été retirées, les tests se poursuivent pour celles qui restent en regardant si deux noeuds sont conditionnellement indépendants. L'ensemble de conditionnement se compose dans un premier temps d'une seule variable parmi les parents de $X$ ou $Y$, puis sa taille augmente incrémentalement jusqu'à ce

qu'une indépendance conditionnelle soit détectée ou que toutes les arêtes liées à $X$ et $Y$ ont été testées. Comme PMIME est asymétrique, l'algorithme teste les deux directions : de $X$ vers $Y$ et de $Y$ vers $X$.
A l'issue de cette première phase, le graphe obtenu est non-dirigé. La règle $R_0$ présentée en section 2.2 est alors mise en oeuvre pour obtenir une pré-orientation du graphe et identifier les *unshielded colliders*. Dans une troisième phase, les ensembles de *Possible-d-sep* sont établis. La première phase est alors réitérée en conditionnant cette fois par les variables se trouvant dans les ensembles de *Possible-d-sep* de $X$ et $Y$ plutôt que par les variables adjacentes.
Enfin, l'algorithme termine par l'utilisation d'un ensemble de règles qui permettent d'orienter

**Algorithme 1 :** *FCI-PMIME*

**Entrées :** $n$ observations de $X^0, \ldots, X^g$, paramètres de PMIME
**Sorties :** $\mathcal{G}$, le graphe estimé
Créer un graphe complet non-dirigé $\mathcal{G} = (V, E)$, avec $V = (V_{t-}, V_t)$ et tel que $X^i_{t-} \circ\!\!\rightarrow X^j_t, \quad \forall X^i_{t-} \in V_{t-}, X^j_t \in V_t$
**pour** *chaque permutation* $(X^i, X^j)$ *de noeuds* $\mathcal{G}$ **faire**
  Calculer la mesure $R$ entre $X^j$ et $X^i$, affecter $R$ à l'arête $V^{i,j}$
Retirer les arêtes $V^{i,j}$ de $V$ si $R \approx 0$
Initialiser $l = 1$, *process* = Vrai, *sep_set* = []
**tant que** *process est Vrai* **faire**
  Définir *process* sur Faux
  **pour** *chaque permutation* $(X^i, X^j) \in V$ **faire**
    $adj\_set = Par(X^i, \mathcal{G}) \backslash X^j$ la liste des prédécesseurs de $X^i$, sans $X^j$
    **si** $Card(adj\_set) \geq l$ **alors**
      **pour** *chaque combinaison* $\mathbf{Z} \in adj\_set$ *de taille* $l$ **faire**
        Calculer $R(X^j \rightarrow X^i \mid \mathbf{Z})$, affecter $R$ à l'arête entre $X^j$ et $X^i$
        **si** $R \approx 0$ **alors**
          ajouter $\mathbf{Z}$ à *sep_set*
    Définir *process* comme Vrai
Retirer les arêtes $V^{i,j}$ de $V$ si $R \approx 0$
Appliquer la règle d'orientation $R_0$ pour détecter les *unshielded colliders*
Identifier les ensembles de *Possible-D-separation*
Répéter la boucle Tant que précédente en itérant sur les membres de *Possible D-separation* entre $X^i$ et $X^j$
Réinitialiser l'orientation du graphe $\mathcal{G}$ : $X^i \circ\!\!-\!\!\circ X^j, \forall X^i, X^j \in V$
Orienter $\mathcal{G}$ en appliquant la règle $R_0$ puis les règles 1 à 4 et 8 à 10 de FCI.

le graphe. L'orientation du graphe est réinitialisée et la règle $R_0$ est de nouveau appliquée. Ensuite, les règles $1, 2, 3$ et $4$, de FCI [Spirtes et al., 2000] sont employées ainsi que les règles $8, 9, 10$ de [Zhang, 2008]. Les règles 5 à 7 de ce dernier traitent des effets cachés et ne sont pas prises en compte dans notre contexte car nous ne nous intéressons qu'aux causes communes cachées. Certaines règles sont aussi adaptées pour les séries temporelles : par exemple, toutes les arêtes telles que $X^i_{t-\tau} \circ\!\!-\!\!\circ X^j_t$ sont orientées $X^i_{t-\tau} \circ\!\!\rightarrow X^j_t$, indiquant simplement que la causalité ne peut se propager dans le passé. Notre méthode, FCI-PMIME est décrit dans l'Algorithme 1.

# 4 Expérimentations

La méthode FCI-PMIME, est évaluée sur des données simulées issues d'un article récent [Assaad et al., 2022b]. Les auteurs ont simulé des graphes causaux à partir de séries temporelles qui forment des structures souvent rencontrées en étude de causalité. Une structure nous intéresse particulièrement, celle nommée $7ts2h$ qui contient des variables cachées. Elle se compose de 7 variables observées et 2 variables cachées, comme représenté sur la Figure 2 (b). Le modèle de génération des données considère des relations linéaires d'auto-corrélation et non-linéaires entre les variables. La structure est simulée 10 fois avec $n = 4000$ observations.
Sur ce jeu de données, FCI-PMIME est comparée à plusieurs autres méthodes issues de l'état de l'art : PWGC [1], VarLiNGAM, DYNOTEARS [2] et PCMCI [3]. Pour chaque algorithme, le décalage maximal considéré est $\tau_{max} = 3$, sélectionné empiriquement. PWGC est utilisé avec un F-test et un seuil de signification $\alpha = 0.03$. VarLiNGAM utilise une pénalisation Lasso paramétrée par le Critère d'Information Bayesien (BIC). La paramétrisation de DYNOTEARS

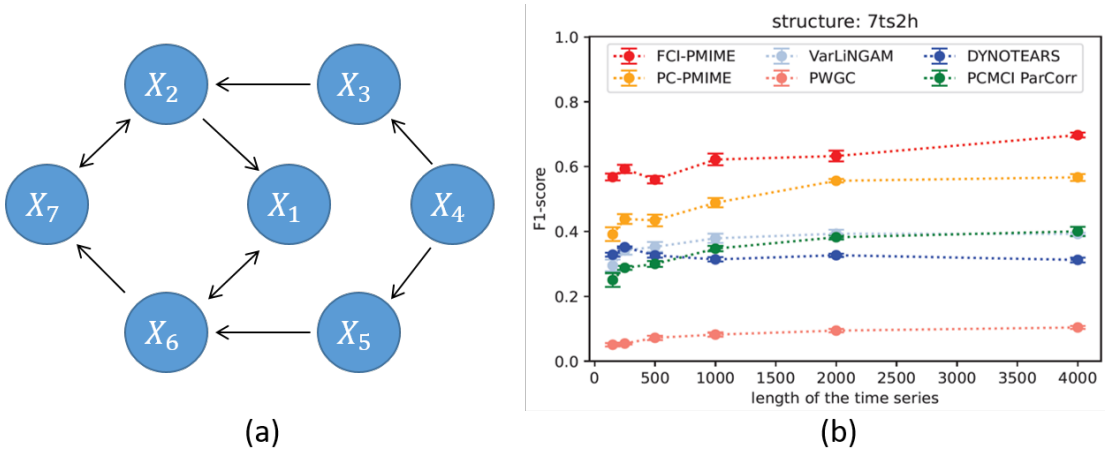



FIGURE 2 – A gauche (a), le graphe simulé du jeu de données $7ts2h$, une double flèche entre deux noeuds indique la présence d'une variable latente. A droite, les résultats des différentes méthodes avec le $F1$-score en ordonnées et la taille des séries temporelles en abscisses.

est celle recommandée par [Pamfil et al., 2020], avec $\lambda_w = 0.05 = \lambda_a$ et le seuil $\tau_w = 0.01$. Pour finir, la mesure utilisée dans PCMCI est la *Partial Correlation* avec un seuil de signification fixé à $\alpha = 0.03$. Le nombre de voisins utilisé pour l'estimation de la CMI dans FCI-PMIME est $k = 0.03 \cdot n$, selon [Frenzel and Pompe, 2007]. Après plusieurs tests, il apparaît que si le critère d'arrêt $A$ est proche de 0 ($A \leq 0.01$), il est trop permissif, à l'inverse, si $A \geq 0.05$, il est trop conservateur. Le critère d'arrêt est donc fixé à $A = 0.03$, suivant les résultats de tests empiriques.

1. `https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.grangercausalitytests.html`.
2. https ://github.com/quantumblacklabs/causalnex
3. `https://github.com/jakobrunge/tigramite`

La métrique employée pour comparer les différentes méthodes est le $F1$-score, où un vrai-positif apparaît lorsqu'une arête dans le graphe estimé est aussi dans le graphe simulé. L'auto-corrélation n'est pas prise en compte dans la mesure du score car toutes les méthodes parviennent à la détecter ce qui fausse grandement le score. Sont alors calculées les moyennes des $F1$-scores obtenus sur 10 graphes simulés pour différentes longueurs de séries temporelles $n = \{125, 250, 500, 1000, 2000, 4000\}$.
Pour chaque longueur de série, la méthode FCI-PMIME obtient des performances supérieures aux autres méthodes. De plus, son score augmente avec la taille des séries temporelles, contrairement à l'algorithme PC-PMIME, par exemple, ce qui montre une certaine consistance. Une légère instabilité est constatée pour des séries de faible taille, expliquée par l'estimation asymptotique de la CMI dans la mesure PMIME. Ainsi, lorsque le nombre d'observation est réduit, les indépendances (conditionnelles) sont plus difficiles à détecter et le graphe inféré comporte souvent un nombre d'arête sur-estimé.

# 5 Conclusion et futurs travaux

Ce papier présente une méthode pour apprendre un graphe causal à partir de données de séries temporelles dans un système complexe où toutes les variables ne sont pas observées. Notre approche, FCI-PMIME, limite le nombre d'hypothèses sur les données et permet d'étudier des relations entre linéaires ou non-linéaires entre variables et décalées dans le temps. La méthode produit des résultats prometteurs sur des données simulées, qui indiquent toutefois qu'une marge de progression est possible. Un test sur des données réelles permettrait d'évaluer la méthode sur des données plus complexes. Dans le domaine de la fabrication, l'objectif consisterait, par exemple, à identifier les facteurs responsables d'une défaillance au sein d'un système complexe sous surveillance.
La plupart des méthodes de découverte causale existantes souffre d'un problème de passage à l'échelle, et la nôtre n'y fait pas exception. En effet, les méthodes peuvent se comporter correctement sur des systèmes composés de peu de variables mais lorsque ce nombre devient important, un manque de stabilité s'observe ainsi qu'une diminution de l'interprétabilité du graphe estimé. Une solution serait par exemple de traiter les dépendances entre groupes de variables plutôt que variable par variable.
Enfin, il s'avère que les tests d'indépendance conditionnelle dans un cadre général souffrent d'une faible puissance statistique [Shah and Peters, 2020], c'est aussi le cas pour la mesure PMIME. Cependant, nous pensons que le critère de sélection et le processus *d'embedding* réalisé dans la méthode permettent d'augmenter cette puissance statistique. De prochains travaux pourraient être d'obtenir des garanties théoriques sur la puissance du test basé sur ce critère de sélection.

# Références

[Arsac et al., 2023] Arsac, A., Lomet, A., and Poli, J.-P. (2023). Causal discovery for time series with constraint-based model and pmime measure. In *When Causal Inference meets Statistical Analysis*.

[Assaad et al., 2022a] Assaad, C. K., Devijver, E., and Gaussier, E. (2022a). Inferring extended summary causal graphs from observational time series. *arXiv preprint arXiv :2205.09422*.

[Assaad et al., 2022b] Assaad, C. K., Devijver, E., and Gaussier, E. (2022b). Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73 :767–819.

[Brodersen et al., 2015] Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1) :247–274.

[Entner and Hoyer, 2010] Entner, D. and Hoyer, P. O. (2010). On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128.

[Frenzel and Pompe, 2007] Frenzel, S. and Pompe, B. (2007). Partial mutual information for coupling analysis of multivariate time series. *Physical review letters*, 99(20) :204101.

[Gerhardus and Runge, 2020] Gerhardus, A. and Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc.

[Hyvärinen et al., 2010] Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5).

[Kugiumtzis, 2013] Kugiumtzis, D. (2013). Direct-coupling information measure from nonuniform embedding. *Physical Review E*, 87(6) :062918.

[Malinsky and Spirtes, 2018] Malinsky, D. and Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, pages 23–47. PMLR.

[Pamfil et al., 2020] Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. (2020). Dynotears : Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR.

[Papana et al., 2013] Papana, A., Kyrtsou, C., Kugiumtzis, D., and Diks, C. (2013). Simulation study of direct causality measures in multivariate time series. *Entropy*, 15(7) :2635–2661.

[Pearl, 2009] Pearl, J. (2009). Causal inference in statistics : An overview. *Statistics surveys*, 3 :96–146.

[Peters et al., 2017] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference : foundations and learning algorithms*. The MIT Press.

[Runge, 2018] Runge, J. (2018). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR.

[Runge et al., 2019a] Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. (2019a). Inferring causation from time series in earth system sciences. *Nature communications*, 10(1) :2553.

[Runge et al., 2019b] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019b). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11) :eaau4996.

[Shah and Peters, 2020] Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure.

[Spirtes et al., 2000] Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

[Vakorin et al., 2009] Vakorin, V. A., Krakovska, O. A., and McIntosh, A. R. (2009). Confounding effects of indirect connections on causality estimation. *Journal of neuroscience methods*, 184(1) :152–160.

[Vowels et al., 2022] Vowels, M. J., Camgoz, N. C., and Bowden, R. (2022). D'ya like dags ? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4) :1–36.

[Zhang, 2008] Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17) :1873–1896.

# Estimation de l'Effet Moyen du Traitement (ATE) en survie causale: Comparaison, Applications et Recommandations Pratiques

Charlotte Voinot [1] & Julie Josse [2] & Bernard Sebastien [3]

[1] *INRIA - SANOFI, France, charlotte.voinot@sanofi.com*
[2] *INRIA, France, julie.josse@inria.fr*
[3] *SANOFI, France, bernard.sebastien@sanofi.com*

**Résumé.** L'estimation de l'effet Moyen du traitement (ATE) constitue l'une des mesures fondamentales en inférence causale visant à évaluer l'impact causal d'un traitement sur une variable de résultat. L'analyse de survie causale se positionne au cœur de cette démarche en cherchant à évaluer l'effet d'un traitement sur la survie des patients au cours du temps. Cependant, malgré l'abondance de littérature en survie causale, l'utilisation des méthodes de Cox demeure prédominante pour évaluer cet effet. Ainsi, l'objectif principal de cette recherche est d'estimer l'effet causal d'un traitement en utilisant des données de survie qui ne proviennent pas forcément d'essais randomisés. Elle entend principalement fournir des recommandations pratiques aux utilisateurs face à la multitude d'informations disponibles ainsi que de mettre en lumière les avantages et les différences par rapport aux approches classiques encore largement utilisées. Pour cela, dans un premier temps, un état de l'art des méthodes de survie causale sera présenté en décrivant les hypothèses d'identifiabilité et les principaux estimateurs, dont les méthodes de pondération, de régression et les approches doublement/triplement robustes. Parmi ces méthodes, on trouve des estimateurs paramétriques, semi-paramétriques et non paramétriques comme les forêts de survie causale. Par la suite, une étude extensive par simulation sera réalisée pour comparer les différents estimateurs, leurs régimes de prédilection et illustrer leurs propriétés théoriques sur des échantillons de taille finie. Pour finir, nous examinerons comment l'ajout de certaines variables dans les modèles de censure, de survie ou de traitement peut impacter la variance des estimateurs.

**Mots-clés.** Survie causale, inférence causale, ATE, données censurées

**Abstract.** Estimating the Average Treatment Effect (ATE) is one of the fundamental measures in causal inference, aimed at assessing the causal impact of a treatment on an outcome variable. Causal survival analysis is at the heart of this approach, seeking to evaluate the effect of a treatment on patient survival over time. However, despite the abundance of literature on causal survival, the use of Cox methods remains predominant for assessing this effect. Thus, the main objective of this research is to estimate the causal effect of a treatment using survival data not necessarily derived from randomized trials. Its main aim is to provide users with practical recommendations in the face of the multitude of information available, and to highlight the advantages and differences compared with the classic approaches still widely used. To this end, we will begin by presenting the state of the art in causal survival methods, describing identifiable assumptions and the main estimators, including weighting, regression and triply/doubly robust approaches. These methods include parametric, semi-parametric and non-parametric estimators such as causal survival forests.An extensive simulation study

will then be carried out to compare the different estimators, their preferred regimes and illustrate their theoretical properties on finite sample sizes. Finally, we will examine how the addition of certain variables in the censoring, survival or treatment models can impact the variance of the estimators.

**Keywords.** Causal survival, causal inference, average treatment effect, censoring

# 1 Background

Causal survival analysis can be seen as the combination of causal analysis and survival analysis: the aim is to assess the causal effect of a treatment on a outcome which is a time until an event occurs. The objective of this article is to provide a comprehensive overview of the different available methods to estimate the (average) effect of a treatment on survival.

## 1.1 Notation

Let's consider a sample of $n$ i.i.d observations that are described by:

- $X_i$: the covariates, $X \in \mathbb{R}^p$
- $A_i$: the binary treatment, $A \in \{0, 1\}$
- $C_i$: the time to censoring, $C \in \mathbb{R}^+$
- $T_i(0)$: the survival time to the event of interest had the patient received control $A_i = 0$
- $T_i(1)$: the survival time to the event of interest had the patient received treatment $A_i = 1$
- $T_i = A_i T_i(1) + (1 - A_i) T_i(0)$, $T \in \mathbb{R}^+$: the observed outcome corresponds to the potential outcome under the assigned treatment; this is known as the consistency identifiability assumption in causal inference
- $\Delta_i = I\{T_i \leq C_i\}$ the status of censoring, where $I\{\cdot\}$ is the indicator
- $\tilde{T}_i = T_i \wedge C_i = \min(T_i, C_i)$, the observed time. When an observation is censured, then its observed time is equal to the censoring time

The observed data can be summarized as a quadruplet $(X_i, A_i, \Delta_i, \tilde{T}_i)$ represented in Table 1.

## 1.2 Treatment effect

In causal inference, the primary goal is to estimate the individual causal effect of the treatment denoted as $\theta_i = T_i(1) - T_i(0)$ (Rubin, 1974). However, this quantity cannot be observed

Table 1: Example of survival data

| | Covariates | | | Treatment | Censoring | Status | Outcomes | | |
|---|---|---|---|---|---|---|---|---|---|
| **ID** | $X_1$ | $X_2$ | $X_3$ | **A** | **C** | $\Delta$ | **T(0)** | **T(1)** | $\tilde{T}$ |
| 1 | 1 | 1,5 | 4 | 1 | ? | 1 | ? | 200 | 200 |
| 2 | 5 | 1 | 2 | 0 | ? | 1 | 100 | ? | 100 |
| 3 | 9 | 0,5 | 3 | 1 | 200 | 0 | ? | ? | 200 |

because at most one outcome can be observed per sample (see Table 1). Furthermore, censoring may also mask outcomes (Turkson et al., 2021). Despite these challenges, certain identifiability assumptions enable us for estimating the average treatment effect (Díaz et al., 2019; Ozenne et al., 2020) (ATE) which is defined as follows:

**Definition 1** (Causal effect: **Average treatment effect** in survival analysis (ATE))**.**

$$\theta = \mathbb{E}\left[y(T(1)) - y(T(0))\right]$$

*where $y(T)$ is some deterministic transformation of the survival time $T$ such as:*

- *$y(T) = I\{T > t\}$ for $t \leq \tau$; then, $E(y(T))$ becomes the survival probability at time $t$.*
- *$y(T) = T \wedge \tau = \min(T, \tau)$ with $\tau$ a fixed time horizon; then, $E(y(T))$ becomes the restricted mean survival time (RMST) at time $\tau$ (Chen and Tsiatis, 2001).*

**Definition 2** (Causal effect: Difference between **survival probabilities**)**.**

$$\theta(t) = E[I\{T(1) > t\} - I\{T(0) > t\})] = S_1\{t) - S_0(t)$$

*with $S_a(t) = P(T(a) > t)$, the probability of surviving at time $t$ when treatment $A = a$.*

**Definition 3** (Causal effect: Difference of **restricted mean survival time (RMST)** between treated and controls)**.**

$$\theta_{RMST}(\tau) = \mathbb{E}\left[T(1) \wedge \tau - T(0) \wedge \tau\right]$$

*Survival probabilities and RMST are linked as follows:*

$$\theta_{RMST}(\tau) = \int_0^\tau \left(S_1(t) - S_0(t)\right) dt$$

RMST can be interpreted as the average survival time from baseline to a pre-specified time $\tau$: a RMST value of 10 days with $\tau = 200$ means that on average the treatment increases the survival time by 10 days at 200 days.

In this paper, we focus on $\theta_{RMST}$ as the estimand of interest. The aim is to construct estimators of this average causal effect while overcoming potential biases due to confounding factors and to right censoring.

## 1.3 Censoring mechanism

Two different assumptions about the censoring mechanism can be considered.

**Assumption 1** (**Independent/ Non informative censoring**).

$$C \perp\!\!\!\perp T(0), T(1), X, A$$

Under Assumption 1, subjects censored at time $t$ are representative of all subjects who remain at risk at time $t$. Therefore, the probability of experiencing an event should be the same for both censored subjects and subjects remaining at risk. It is as if the censored subjects were randomly selected from all subjects.

**Assumption 2** (**Conditionally independent censoring**).

$$C \perp\!\!\!\perp T(0), T(1) | X, A$$

Under Assumption 2, within subgroups represented by $X = x$, subjects censored at time $t$ are representative of all subjects in their subgroup who remain at risk at time $t$. It is as if the censored subjects were randomly selected inside each subgroup. This assumption is also referred to as dependent censoring.

But another assumption for identifiability of RMST is required under Assumption 2: we need to assume that all subjects have a positive probability to remain uncensored at their failure time.

**Assumption 3** (**Positivity / Overlap for censoring** ).

$$pr(C > t \mid X = x, A = a) > 0, \quad \textit{for the identifiability of RMST: } t \leq \tau.$$

If for a time $t$, $\mathbb{P}(C > t \mid X = x, A = a) = 0$, then this excludes that we have any observed outcomes after time $t$. For example, if we consider a clinical trial with administrative censoring after one year of study, then the probability of remaining uncensored after one year is zero. In that case, the potential outcomes $T(0)$ and $T(1)$ are not observed at all after $t = 1$ year. One can consider lowering the threshold time $\tau$ such that each subject has a probability of remaining uncensored at their restricted time.

# 2 Causal survival analysis with a Randomized Control Trial

Randomized clinical trials (RCTs) are the gold standard for establishing the effect of a treatment on an outcome, because treatment allocation is under control, which ensures (asymptotically) the balance of covariates between treated and controls, and thus avoids problems of confounding between covariables and treatment.

The core assumption in a RCT is the random assignment of the treatment (Rubin, 1974).

**Assumption 4** (**Random treatment assignment**).

$$A \perp\!\!\!\perp (T(0), T(1), X)$$

Assumption 4 implies that the treatment is given at random and is independent of both the potential outcomes and the covariates.

**Indentifiability.** Under Assumptions 4 (random treatment assignment) and 1 (independent censoring), the RMST can be identified as follows:

$$\begin{aligned}
\theta_{RMST} = \mathbb{E}[T(1) \wedge \tau - T(0) \wedge \tau] &= \int_0^\tau \mathbb{E}[I\{T(1) > t\} - I\{T(0) > t\}]dt && \text{(Independent censoring)} \\
&= \int_0^\tau \mathbb{E}[I\{T(1) > t | A = 1\} - \mathbb{E}[I\{T(0) > t | A = 0\}dt && \text{(Random treatment assignment)} \\
&= \int_0^\tau \mathbb{E}[I\{T > t | A = 1\} - \mathbb{E}[I\{T > t | A = 0\}]dt && \text{(Consistency)} \\
&= \int_0^\tau S(t|A = 1) - S(t|A = 0)dt
\end{aligned} \tag{1}$$

where $S(t|A = a)$ is the survival function of the population with treatment $A = a$.

Under Assumptions 4 (random treatment assignment) and 2 (conditionally independent censoring), the RMST can be identified as follows:

$$\begin{aligned}
\theta_{RMST} &= \int_0^\tau \mathbb{E}[I\{T > t | A = 1, X\}] - \mathbb{E}[I\{T > t | A = 0, X\}]dt && \text{(Consistency)} \\
&= \int_0^\tau \mathbb{E}\left[\frac{I\{T > t | A = 1\} * \Delta}{S_c(t|X, A = 1)} - \frac{I\{T > t | A = 0\} * \Delta}{S_c(t|X, A = 0)}\right] dt && \text{(Conditionally independent censoring)}
\end{aligned} \tag{2}$$

where $S_c(t|X, A = a)$ is the survival function of remain uncensored given the covariate $X_i$

## 2.1 Estimation under independent censoring

### 2.1.1 Non adjusted Kaplan meier estimator

**Definition 4** (Unadjusted kaplan meier estimator)**.**

$$\hat{S}_{KM}(t \mid a) == \prod_{j=1, t_j <= t} \left(1 - \frac{\sum_i I\left\{\tilde{T}_i = t_j, \Delta_i = 1, A_i = a\right\}}{\sum_i I\left\{\tilde{T}_i \geq t_j, A_i = a\right\}}\right)$$

Unadjusted Kaplan meier which maximizes the likelihood of the observations is a uniformly consistent non parametric estimator for estimating the survival function (Gill, 1983) and (Kaplan and Meier, 1958).

The corresponding RMST is obtained in integrating from 0 to $\tau$ the difference between non adjusted kaplan meier estimator of the treated and controls (1).

## 2.2 Estimation under conditional censoring

Under Assumptions 4 (random treatment assignment), 2 (conditional censoring) and 3 (censoring positivity), the unadjusted KM estimator overestimates the real survival probabilities (Willems et al., 2018). Thus, correction for the presence of dependent censoring is important in order to obtain a good estimator. Under these assumptions, the adjusted IPCW (inverse probability of censoring weighting) Kaplan meier estimator (Robins and Rotnitzky, 1992; Robins and Finkelstein, 2000) can be used to estimate the causal treatment effect.

### 2.2.1 (IPCW) adjusted Kaplan meier estimator

**Definition 5** (IPCW adjusted kaplan meier estimator)**.**

$$\hat{S}_{IPCW-KM}(t \mid a) = \prod_{j=1, t_j <= t} \left(1 - \frac{\sum_i \hat{w}_i(t_j, X_i) * I\left\{T_i = t_j, C_i \geq t_j, A_i = a\right\}}{\sum_i \hat{w}_i(t_j, X_i) * I\left\{T_i \geq t_j, C_i \geq t_j, A_i = a\right\}}\right)$$

- $\hat{w}_i(t, X_i) = \frac{1}{\hat{S}_c(t|X_i, A_i)}$ *is the inverse of probability of remain uncensored given* $X_i$*.*
- $\hat{S}_c(t|X_i, A_i)$ *is based on the fit of semi-parametric or parametric model for censoring (for example a Cox model) with* $X_i$ *and* $A_i$ *the covariates.*

This estimator gives extra weight to subjects who are not censored. At every time point $t$, each subject $i$ is given a weight which is inversely proportional to the estimated probability of having remained uncensored until time $t$.

In the same way than before, the corresponding RMST is obtained in integrating from 0 to $\tau$ the difference between adjusted kaplan meier estimator of the treated and controls (2).

# 3 Causal survival analysis with an observational study

In the context of observational study, Assumption 4 (randomized treatment assignment) is no longer verified. Some additional assumptions are required to identify $\theta_{RMST}$. These assumptions are classical for causal inference with observational data:

**Assumption 5** (**Conditional exchangeability / Uncounfoundedness** ).

$$A \perp\!\!\!\perp (T(0), T(1))|X$$

*with $X$ the set of covariates that are related both to treatment's assignment and outcomes.*

Under Assumption 5, the treatment assignment is randomly assigned conditionally on the covariates $X$. It is as if the treatment for all subjects were randomly selected inside each subgroup.

**Assumption 6** (**Positivity / Overlap for treatment**).

$$1 > P(A = a \mid X = x) > 0$$

**Identifiability.** Under assumption 5 (Uncounfoundedness) and 1 (Independent censoring), the RMST can be identified as follows:

$$\begin{aligned}
\theta &= \mathbb{E}[T(1) \wedge \tau - T(0) \wedge \tau] \\
&= \int_0^\tau \mathbb{E}\left[ I\{T(1) > t\} * \frac{A}{e(X)} - I\{T(0) > t\} * \frac{1-A}{1-e(X)} \right] dt \text{ (Identifiability of the IPTW)} \\
&= \int_0^\tau \mathbb{E}\left[ I\{T > t \mid A = 1\} * \frac{A}{e(X)} - I\{T > t \mid A = 0\} * \frac{1-A}{1-e(X)} \right] dt \text{ (By consistency)}
\end{aligned} \tag{3}$$

Under assumption 5 (Uncounfoundedness) and 2 (conditionally independent censoring), the RMST can be identified as follows:

$$\begin{aligned}
\theta &= \mathbb{E}\left[ (T \wedge \tau) \left( \frac{A}{e(X)} - \frac{1-A}{1-e(X)} \right) \right] \text{ (Identifiability of the IPTW)} \\
&= \mathbb{E}\left[ \mathbb{E}[(T \wedge \tau) \mid A, X] \left( \frac{A}{e(X)} - \frac{1-A}{1-e(X)} \right) \right] \\
&= \mathbb{E}\left[ \frac{\widetilde{T} \wedge \tau \cdot \Delta^\tau}{S_C(\widetilde{T} \wedge \tau \mid A, X)} \left( \frac{A}{e(X)} - \frac{1-A}{1-e(X)} \right) \right]
\end{aligned} \tag{4}$$

Under the same assumptions, it can be identified also as g-formula:

$$\begin{aligned}
\theta &= \mathbb{E}\left[ T(1) \wedge \tau - T(0) \wedge \tau \right] \\
&= \mathbb{E}\left[ \mathbb{E}\left[ T(1) \wedge \tau \mid X, A = 1 \right] - \mathbb{E}\left[ T(0) \wedge \tau \mid X = X, A = 0 \right] \right] \text{ (Uncounfoundedness)} \\
&= \mathbb{E}\left[ \mathbb{E}\left[ T \wedge \tau \mid X, A = 1 \right] - \mathbb{E}\left[ T \wedge \tau \mid X, A = 0 \right] \right] \text{ (Consistency)}
\end{aligned} \tag{5}$$

## 3.1 Estimation under independent censoring

### 3.1.1 IPTW Kaplan meier estimator

Under Uncounfoundedness and independent censoring (Assumptions 5 and 1), the Kaplan meier estimator has to include a weighting term to take into account that the treated and control groups are unbalanced. This weighted estimator is called the inverse probability of treatment weighted Kaplan meier estimator (IPTW-KM) (Xie and Liu, 2005).

Based on the identifiability (3), the IPTW KM is defined as:

**Definition 6** (Adjusted IPTW kaplan meier estimator)**.**

$$\hat{S}_{IPTW-KM}(t \mid a) = \prod_{j=1, t_j <= t} \left(1 - \frac{\sum_i \hat{w}_i * I\left\{T_i = t_j, C_i \geq t_j, A_i = a\right\}}{\sum_i \hat{w}_i * I\left\{T_i \geq t_j, C_i \geq t_j, A_i = a\right\}}\right)$$

*with* $\hat{w}_i = \frac{A_i}{\hat{e}(X_i)} + \frac{1-A_i}{1-\hat{e}(X_i)}$ *the inverse of the propensity score.*

In the exact same way than before, the corresponding RMST is obtained in integrating from 0 to $\tau$ the difference between IPTW adjusted kaplan meier estimator of the treated and controls (3).

## 3.2 Estimation under conditionally independent censoring

### 3.2.1 Inverse probability of weighting estimation (IPTW-IPCW)

When the independent censoring assumption is not verified, the IPTW-IPCW Kaplan meier estimator can be used to estimate the causal treatment effect . The IPTW-IPCW KM estimator is a combination of both estimators (Anstrom and Tsiatis, 2004): IPTW to overcome that the treatment allocation is not random and IPCW to overcome the dependent censoring.

Based on the identifiability (4), the IPTW-IPCW is defined as:

**Definition 7** (IPTW-IPCW estimator)**.**

$$\widehat{\theta}_{\text{IPTW-IPCW}}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i^{\tau} \cdot \tilde{T}_i \wedge \tau}{\hat{S}_C\left(\widetilde{T} \wedge \tau \mid A_i, X_i\right)} \left(\frac{A_i}{\hat{e}\left(X_i\right)} - \frac{1 - A_i}{1 - \hat{e}\left(X_i\right)}\right)$$

*where* $\hat{S}_C\left(\widetilde{T} \wedge \tau \mid A_i, X_i\right)$ *is a semi parametric (or parametric) methodology to estimate the survival function of remain uncensored.*

It enables a balance between treatment and control groups and between censored and uncensored individuals.

This RMST estimator can be obtained at first in estimating the survival function of

remain uncensored in using semi-parametric or parametric model. Then, the weight for each observation can be computed as:

$$w_i = \frac{\Delta_i^\tau}{\hat{S}_C\left(\widetilde{T} \wedge \tau \mid A_i, X_i\right)} * \left(\frac{A_i}{\hat{e}\left(X_i\right)} - \frac{1 - A_i}{1 - \hat{e}(X_i)}\right)$$

An adjusted kaplan estimator (weighted by the previous $w_i$) can be fitted for $A = 1$ and $A = 0$ (see definition 8).

**Definition 8** (Adjusted IPTW-IPCW kaplan meier estimator)**.**

$$\hat{S}_{IPTW-IPCW-KM}(t \mid a) = \prod_{j=1, t_j <= t} \left(1 - \frac{\sum_i \hat{w}_i(t, X_i) * I\left\{T_i = t_j, C_i \geq t_j, A_i = a\right\}}{\sum_i \hat{w}_i(t, X_i) * I\left\{T_i \geq t_j, C_i \geq t_j, A_i = a\right\}}\right)$$

*with* $\hat{w}_i(t, X_i) = \frac{1}{\hat{S}_C\left(\widetilde{T} \wedge \tau | A_i, X_i\right)} * \left(\frac{A_i}{\hat{e}(X_i)} + \frac{1 - A_i}{1 - \hat{e}(X_i)}\right)$ *the corresponding weight including the inverse of the propensity score and the inverse probability of remain uncensored given the covariates.*

Then, the corresponding RMST is the integral of the difference between the survival curve with $A = 1$ and the $A = 0$.

### 3.2.2 G-formula plug-in estimator

Another possible estimator under assumption 5 and 2 is the G-formula plug-in estimator.

It is an alternative of IPCW in leveraging the regression formulation. Instead of fitting a model for the censored mechanism and a model for the probability of being treated, the corresponding estimators fit a model of the conditional outcome mean. Applying these models to the each treatment arm, and then marginalizing over the empirical covariates distributions of the target population, gives the corresponding expected outcome (Robins, 1986). Based on the g-formula identifiability (5), this outcome model based estimator is defined as:

**Definition 9** (G-formula plug-in estimator)**.**

$$\widehat{\theta}_{g\text{-}formula}\left(\tau\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{F}\left(X_i, 1\right) - \hat{F}\left(X_i, 0\right)\right)$$

*with* $F(x, a) \triangleq \mathbb{E}[T \wedge \tau \mid X = x, A = a]$*. It can be estimated in using semi-parametric or parametric methods.*

Generally, $F(x, a)$ estimator is based on the estimation of the conditional survival function. It can be obtained in fitting one semi-parametric (or parametric) model (i.e. Cox model) by treatment on the corresponding observations and in predicting the results for the all observations. Then, the RMST is computed by the integral of the difference between the predicted conditional survival curve with A=1 and A=0.

# References


Anstrom, K. J. and A. A. Tsiatis (2004, 05). Utilizing Propensity Scores to Estimate Causal Treatment Effects with Censored Time-Lagged Data. *Biometrics 57*(4), 1207–1218.

Chen, P.-Y. and A. A. Tsiatis (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics 57*(4), 1030–1038.

Díaz, I., E. Colantuoni, D. F. Hanley, and M. Rosenblum (2019, July). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime data analysis 25*(3), 439—468.

Gill, R. (1983). Large Sample Behaviour of the Product-Limit Estimator on the Whole Line. *The Annals of Statistics 11*(1), 49 – 58.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Ozenne, B. M. H., T. H. Scheike, L. Stærk, and T. A. Gerds (2020). On the estimation of average treatment effects with right-censored time to event outcome and competing risks. *Biometrical Journal 62*(3), 751–763.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9), 1393–1512.

Robins, J. M. and D. M. Finkelstein (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics 56*(3), 779–788.

Robins, J. M. and A. Rotnitzky (1992). *Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers*, pp. 297–331. Boston, MA: Birkhäuser Boston.

Rubin, D. B. (1974, October). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Turkson, A., F. ayiah mensah, and V. Nimoh (2021, 09). Handling censoring and censored data in survival analysis: A standalone systematic literature review. *International Journal of Mathematics and Mathematical Sciences 2021*, 1–16.

Willems, S., A. Schat, M. van Noorden, and M. Fiocco (2018). Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical Methods in Medical Research 27*(2), 323–335. PMID: 26988930.

Xie, J. and C. Liu (2005). Adjusted kaplan–meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine 24*(20), 3089–3110.

# Généralisation du rapport de risques en utilisant des données observationnelles

Ahmed Boughdiri [1], Julie Josse [2] & Erwan Scornet [3]

[1] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, ahmed.boughdiri@inria.fr*
[2] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, julie.josse@inria.fr*
[3] *Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France, erwan.scornet@sorbonne-universite.fr*

**Résumé.** Les essais contrôlés randomisés (ECR) sont actuellement la référence pour mesurer empiriquement un effet causal d'une intervention donnée sur une réponse. En pratique, presque tous les nouveaux médicaments sont autorisés grâce à de tels essais. Mais récemment, des préoccupations ont été soulevées concernant la portée limitée des ECR : critères d'éligibilité stricts, conformité non réaliste dans le monde réel, cadre temporel court, taille d'échantillon limitée, etc. De telles limitations menacent la validité externe des études ECR dans d'autres situations ou populations. L'utilisation de données non randomisées complémentaires, appelées observationnelles, est prometteuse. La transportabilité permet de généraliser les réponses de l'essai vers une population cible d'intérêt, dont la distribution diffère par rapport à la population de référence. Cependant, certaines préoccupations ont été soulevées, en particulier lorsque la population cible est très différente de la cohorte ECR. De plus, ces méthodes sont encore à un stade de prototype, avec un grand écart entre la théorie et la pratique, et la confrontation avec les données soulève de nombreux défis méthodologiques. Des développements et validations théoriques, méthodologiques et appliqués sont nécessaires pour mieux comprendre et exploiter l'équilibre entre rapidité et sécurité, et pour concevoir les futurs ECR. La différence de risques (DR) est une mesure causale souvent utilisée pour caractériser l'effet d'un médicament. Différentes méthodes (systèmes de pondération, modélisation des réponses ou approches doublement robustes) ont été proposées et étudiées pour estimer la DR. Cependant, selon le type de réponse, d'autres mesures comme le rapport de risques (RR) peuvent être plus appropriées. Nous visons d'abord à étendre les différentes classes d'estimateurs au rapport de risques ainsi que leurs propriétés théoriques, en particulier pour des échantillons finis. Nous proposons ensuite différents estimateurs de généralisation pour le rapport de risques.

**Mots-clés.** Inférence causale, Essais Contrôlés Randomisés, Données observationnelles, Généralisation

**Abstract.** Randomized Controlled Trials (RCT) are the current gold-standard to empirically measure a causal effect of a given intervention on an outcome. In practice, almost all new drugs are authorized through such trials. But recently, concerns have been raised on the limited scope of RCTs: stringent eligibility criteria, unrealistic real-world compliance, short timeframe, limited sample size, etc. Such limitations threaten the external validity of RCT

studies to other situations or populations. The usage of complementary non-randomized data, referred to as observational brings promises as additional sources of evidence. Transportability allows to generalize the trial findings toward a target population of interest, potentially subject to a covariate distributional shift. However, some concerns have been raised, especially when the target population is very different from the RCT cohort. In addition, such methods are still in a prototype stage, with a wide gap between theory and practice and the confrontation with data give rise to many methodological challenges. Theoretical, methodological and applied developments and validations are needed to better understand and leverage the speed-safety balance and to design future RCTs. The Risk Difference (RD) is a causal measure often used to characterized the effect of a drug. Different methods (weighting schemes, outcome modeling, or doubly-robust approaches) have been proposed and studied to estimate the RD. However, depending on the type of output, other measures as the Risk Ratio (RR) may be more appropriate. We aim at first extending the different classes of estimators to the Risk Ratio and derive theoretical guarantees, in particular finite-sample guarantees. We then propose different generalizing estimators for the risk ratio.

**Keywords.** Causal inference, Randomized Controlled Trials, Observational data, Generalization

# 1 Cadre et notation

Le cadre des réponses potentiels, initié par Neyman en 1923 et largement diffusé par Rubin (Rubin, 1974) dans les années 1970, joue un rôle crucial dans la compréhension des effets des traitements dans les Essais Contrôlés Randomisés (ECR). Considérons un individu $i$ ; soit $T_i \in \{0, 1\}$ la variable aléatoire qui désigne l'attribution du traitement, avec 1 indiquant la prise du traitement et 0 l'absence de traitement. Les réponses potentiels, notés $Y_i^{(1)}$ pour le cas où le traitement est administré et $Y_i^{(0)}$ pour le cas contraire, représentent respectivement les issues possibles avec et sans le traitement.

La réponse observé $Y_i$ est défini comme suit :

$$Y_i = Y_i^{(T_i)} = \begin{cases} Y_i^{(0)}, & \text{si } T_i = 0 \\ Y_i^{(1)}, & \text{si } T_i = 1 \end{cases}$$

L'effet du traitement individuel, $\delta_i = Y_i^{(1)} - Y_i^{(0)}$, quantifie l'efficacité du traitement. Cependant, seul l'un des réponses potentiels peut être observé pour chaque individu. On suppose que l'on observe $n$ individus, indexés par $i \in \mathcal{I}$. Pour chaque unité $i$ et pour chaque niveau de traitement, il existe des réponses potentiels correspondants $Y_i^{(0)}$ et $Y_i^{(1)}$. Pour $t \in \{0, 1\}$, nous définissons les variables aléatoires $Y^{(t)}$ et $T$ qui associent respectivement à $i \in \mathcal{I}$ $Y_i^{(t)}$ et $T_i$. On definit alors l'Effet Moyen du Traitement ou la différence de risques, comme l'esperance de l'effet du traitement sur la population $\mathcal{I}$.

**Définition 1** (**Effet Moyen du Traitement/différence de risques (RD)**).

$$\tau_{RD} := \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]$$

Ce dernier peut être interprêté comme la différence entre la réponse moyenne si toute la population avait reçu le traitement et la réponse moyenne si la population n'avait reçu le traitement. De façon similaire, nous définissons le rapport de risques comme le rapport des espérances des réponses potentielles.

**Définition 2** (**Rapport de risques (RR)**)**.**

$$\tau_{RR} := \frac{\mathbb{E}\left[Y^{(1)}\right]}{\mathbb{E}\left[Y^{(0)}\right]}$$

Le rapport de risques fournit une mesure de la probabilité relative qu'ont les individus du groupe exposé de subir l'événement, par rapport à ceux du groupe non exposé. Un rapport de risques de 1 indique qu'il n'y a pas d'association, un rapport de risques supérieur à 1 suggère un risque accru avec l'exposition, et un rapport de risques inférieur à 1 suggère un risque diminué avec l'exposition.

Afin d'estimer ces quantités causales et garantir l'interprétation valide de la différence de risques (RD) et du rapport de risques (RR), il est essentiel de prendre en considération les hypothèses fondamentales de l'inférence causale.

- **Hypothèse de Consistance (Consistency)** :

  $$\forall (t, i) \in \{0, 1\} \times \mathcal{I}, T = t \Rightarrow Y_i = Y_i^{(t)}$$

  En d'autres termes, l'assignation du traitement ne modifie pas de manière intrinsèque la réponse potentielle de l'individu, assurant ainsi que la réponse observée sous le traitement est la même que celle qui aurait été observée si, dans toutes les circonstances, l'individu avait été traité de la même manière.

- **Hypothèse d'Absence d'Interférence (No Interference)** :

  $$\forall i \in \mathcal{I}, Y_i(t_1, ..., t_i, ..., t_n) = Y_i(t_i)$$

  L'hypothèse d'absence d'interférence postule que la réponse potentielle d'un individu $i$, $Y_i^{(t)}$, n'est pas affectée par le traitement administré à un autre individu $j$.

Dans le cadre d'un traitement binaire, ces deux hypothèses peuvent se traduire par l'hypothèse **SUTVA (Stable Unit Treatment Value Assumption)** :

$$Y = TY^{(1)} + (1 - T)Y^{(0)}$$

- **Hypothèse de Non-Confondance (Unconfoundedness)** :

  $$T \perp\!\!\!\perp (Y(0), Y(1))|X$$

  Une fois que l'on contrôle les variables de confusion potentielles, la décision de traiter ou non un individu ne devrait pas être liée aux réponses potentiels. Cette condition est essentielle pour établir une relation causale, car elle permet de comparer les groupes de traitement et de contrôle comme s'ils avaient été randomisés, en supposant que toutes les variables de confusion pertinentes ont été correctement mesurées et ajustées.

- **Hypothèse de Positivité (positivity)** :

$$\exists \eta \in ]0;1[, \forall x, \quad \eta \leq e(x) \leq 1-\eta,$$

où le score de propensition $e(.)$ est la probabilité qu'un individu soit traité sachant ses caractéristiques :

$$e(x) = \mathbb{P}\left[T = 1 | X = x\right].$$

Dit autrement, cela signifie qu'il n'existe pas de sous-populations avec des probabilités nulles ou quasi nulles de recevoir un traitement spécifique.

- **Hypothèse non-négativité (non-negativity)** :

$$\forall i, j \quad Y_i Y_j \geq 0$$

Cette dernière hypothèse contraint les réponses de tous les individus à avoir le même signe, ce qui permet ainsi au rapport de risques d'être bien defini.

# 2 Estimation du rapport de risques avec des données observationnelles

Sous ces hypothèses, le rapport de risques peut être identifié à partir des differentes approches suivantes issues des données observationnelles :

1. Approche de la repondération:

$$\tau_{RR} = \frac{\mathbb{E}\left[\frac{TY}{e(X)}\right]}{\mathbb{E}\left[\frac{(1-T)Y}{1-e(X)}\right]}$$

2. Approche de la régression:

$$\tau_{RR} = \frac{\mathbb{E}\left[\mu_1(X)\right]}{\mathbb{E}\left[\mu_0(X)\right]}$$

À partir de ces differentes approches et de façon similaire au cas de l'effet moyen du traitement, nous pouvons construire des estimateurs du rapport de risques. En effet, de l'approche de la repondération découle le Ratio Inverse Propensity Weighting (R-IPW).

**Définition 3 (Ratio Inverse Propensity Weighting (R-IPW)).**

$$\tau_{R-IPW} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}},$$

*avec* $\hat{e}(X) = \hat{\mathbb{P}}\left[T = 1 | X\right]$.

Cet estimateur s'inpire de l'estimateur IPW introduit par (Hirano, Imbens, and Ridder, 2003). Ainsi comme l'IPW, le R-IPW augmente le poids des observations traitées ayant un faible score de propension (et inversement) pour équilibrer les deux groupes, traités et non traités, en fonction de leurs covariables.

De l'approche de la regression on peut également contruire un estimateur similaire à la G-formula dans le cas de la différence.

**Définition 4** (**Ratio G-formula**).

$$\tau_{R-G} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i)}{\sum_{i=1}^{n} \hat{\mu}_0(X_i)},$$

*où* $\mu_a(X) = \mathbb{E}\left[Y|T=a,X\right]$.

Cet estimateur, tout comme pour la différence, estime les deux réponses potentielles pour chaque individu, les moyennes ensuite, puis calcule le ratio. En plus de ces deux approches, comme detaillé dans (Laan and Rose, 2011), (Kennedy, 2015) et (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey, 2017) dans le cadre de la différence pour une estimation non-paramétrique, on peut également construire un estimateur dit double robuste.

**Définition 5** (**Ratio AIPW**).

$$\tau_{R-AIPW} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}},$$

*où* $\hat{e}(X) = \hat{\mathbb{P}}\left[T=1|X\right]$ *and* $\hat{\mu}_a(X) = \mathbb{E}\left[Y|T=a,X\right]$.

Nous démontrons la convergence en loi vers une distribution Gaussienne pour chacun de ces estimateurs, dont les variances seront explicitées.

# 3 Generalisation du rapport de risque

La généralisation a pour but d'élargir la portée des conclusions causales au-delà des conditions initiales de l'étude menée par essai de contrôle randomisé, permettant ainsi une application plus large et plus pertinente des réponses de recherche. Dans notre cas, on cherche à généraliser une étude ECR mené sur une population peu représentative que l'on note $\mathcal{R} = \{1, \ldots, n\}$ vers une population plus représentative dite cible correspondant aux données observations $\mathcal{O} = \{n+1, \ldots, n+m\}$. On suppose de plus que seules les covariables des données observationnelles sont connues. On note $p_{\mathrm{R}}(x)$ (resp $p_{\mathrm{T}}(x)$) la densité des covariables dans l'ECR (resp les données observations). On suppose alors l'hypothèses suivante.

- **Hypothèse d'Inclusion sur les support (Overlap)** :

$$\text{supp}(P_{\text{T}}(X)) \subset \text{supp}(P_{\text{R}}(X)).$$

Cela implique la nécessité d'une représentation adéquate de la distribution des covariables au sein de l'échantillon de l'essai comparativement à la population cible, signifiant que chaque membre de la population cible doit avoir une probabilité non nulle d'être choisi pour l'essai. Il est également crucial d'assurer une répartition équilibrée des covariables entre les différents groupes de traitement au sein de l'échantillon de l'essai. L'attribution du traitement dans l'essai doit respecter le principe de positivité, reprenant ainsi l'hypothèse de positivité évoquée précédemment.

Une dernière hypothèse concernant le comportement des réponses entre la population cible et l'ECR est alors nécessaire afin de correctement estimer les differentes quantités causales sur la population cible. Deux approches sont alors possibles.

1. **Hypothèse de Transportabilité (Transportability)** :

$$\forall t \in \{0,1\} \quad \mathbb{E}_{\text{R}}[Y(t) \mid X] = \mathbb{E}_{\text{T}}[Y(t) \mid X]$$

Cette hypothèse traduit le fait que le comportement des réponses d'un individu ne varie pas selon qu'il se trouve dans l'ECR ou dans la population cible.

2. **Hypothèse de Transportabilité de l'effets de traitement individuel (Transportability of the CATE)** :

$$\mathbb{E}_{\text{R}}[Y(1) - Y(0) \mid X] = \mathbb{E}_{\text{T}}[Y(1) - Y(0) \mid X]$$

Cette denière hypothèse un peu plus faible que la Transportabilité, affirme que seul l'effet de traitement pour un individu ne varie pas selon qu'il se trouve dans l'ECR ou dans la population cible.

Sous l'hypothèse 1 (Colnet, Josse, Varoquaux, and Scornet, 2022) montre que l'estimateur plug-in G-formula suivant converge en norme $L^1$ vers la différence de risques au sein de la population cible.

**Définition 6** (**Plug-in G-formula**)**.**

$$\hat{\tau}_{g,n,m} = \frac{1}{m} \sum_{i=n+1}^{n+m} \left(\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i)\right),$$

*avec* $\hat{\mu}_t(x) = \mathbb{E}_{R}[Y \mid X = x, T = t]$.

On montre alors une formule similaire pour le rapport de risque avec le plug-in R-G-formula.

**Définition 7** (**Plug-in R-G-formula**).

$$\hat{\tau}_{r-g,n,m} = \frac{\sum_{i=n+1}^{n+m} \hat{\mu}_{1,n}(X_i)}{\sum_{i=n+1}^{n+m} \hat{\mu}_{0,n}(X_i)}$$

*avec* $\hat{\mu}_t(x) = \mathbb{E}_{R}[Y \mid X = x, T = t]$.

Sous l'hypothèse 2, (Colnet et al., 2022) montre que l'estimateur IPSW suivant converge en norme $L^1$ vers la différence de risques au sein de la population cible.

**Définition 8** (**IPSW: inverse propensity sampling weighting**).

$$\hat{\tau}_{IPSW,n,m} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{\hat{p}_{T}(X_i)}{\hat{p}_{R}(X_i)} \left( \frac{TY}{e_{R}(x)} - \frac{(1-T)Y}{1-e_{R}(x)} \right).$$

On montre alors une formule similaire pour le rapport de risque avec le R-IPSW:

**Définition 9** (**IPSW: inverse propensity sampling weighting**).

$$\hat{\tau}_{R-IPSW,n,m} = \frac{\sum_{i \in \mathcal{R}} \frac{\hat{p}_{T}(X_i)}{\hat{p}_{R}(X_i)} \frac{TY}{e_{R}(x)}}{\sum_{i \in \mathcal{R}} \frac{\hat{p}_{T}(X_i)}{\hat{p}_{R}(X_i)} \frac{(1-T)Y}{1-e_{R}(x)}}.$$

# References


Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017, 01). Double/debiased/neyman machine learning of treatment effects. *American Economic Review 107*.

Colnet, B., J. Josse, G. Varoquaux, and E. Scornet (2022, 11). Causal effect on a target population: A sensitivity analysis to handle missing covariates. *Journal of Causal Inference 10*, 372–414.

Hirano, K., G. Imbens, and G. Ridder (2003, 02). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–1189.

Kennedy, E. (2015, 10). Semiparametric theory and empirical processes in causal inference.

Laan, M. and S. Rose (2011, 01). *Targeted Learning: Causal Inference for Observational and Experimental Data*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology 66*(5), 688–701.

# Federated Causal Inferences: Estimating the Average Treatment Effect in a Decentralized Setting

Rémi Khellaf [1] & Aurélien Bellet [2]& Julie Josse [3]

[1] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France, remi.khellaf@inria.fr*
[2] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France aurelien.bellet@inria.fr*
[3] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France julie.josse@inria.fr*

**Summary** This article aims at discussing the practical challenges of conducting causal inference analyses with decentralized data. Federated learning brings tools to deal with distributed datasets where privacy are important stakes. Federated causal inference then appears to show particular promise in domains where the data are siloed, while allowing researches to harness the power of large-scale data to answer causal questions. We focus on estimating the Average Treatment Effect (ATE) within a federated framework.

**Résumé** Cet article vise à discuter des défis pratiques liés à la réalisation d'analyses d'inférence causale avec des données décentralisées. L'apprentissage fédéré apporte des outils pour traiter les ensembles de données distribués où la confidentialité est une préoccupation importante. L'inférence causale fédérée semble alors prometteuse dans les domaines où les données sont cloisonnées, tout en permettant aux chercheurs d'exploiter la puissance des données à grande échelle pour répondre à des questions causales. Nous nous concentrons sur l'estimation de l'Effet Traitement Moyen (ATE) dans un cadre fédéré.

**Keywords** Federated learning; Causal inference; regression; privacy; linear model; meta-analysis; average treatment effect; G-Formula; linear models; hospital data

**Mots-clés** Apprentissage fédéré ; Inférence causale ; régression ; modèle linéaire ; méta ; analyse ; effet de traitement moyen ; G ; Formula ; modèles linéaires ; effets de traitement hétérogènes ; données hospitalières ; données réelles ; confidentialité

**Abstract.** Federated causal inference is a burgeoning field at the intersection of machine learning and causal inference, aimed at harnessing data from multiple distributed sources to infer causal relationships while preserving privacy and security. Its applications are the most promising in fields where there is a substantial gain at keeping the data decentralized for logistic or privacy reasons, which is particularly the case of hospitals holding patients data.

In this paper, we explore the challenges of federated estimation of the Average Treatment Effect (ATE) by defining and studying several suitable estimators derived from the plugin G-Formula estimator. We provide a comprehensive comparison of the variances of these estimators. We also compare the sample size and inverse variance aggregation methods for the meta estimators in linear models.

# 1 Introduction

## 1.1 Causal inference in observational studies

Let $X$ denote the $p$-dimensional vector of covariates that belongs to a covariate space $\mathcal{X} \subset \mathbb{R}^p$, $W \in \mathcal{W} = \{0, 1\}$ denote the binary treatment, and $Y \in \mathbb{R}$ denote the outcome of interest. We consider the potential outcomes framework, where for $w \in \mathcal{W}$, $Y(w)$ is the outcome had the subject received treatment $w$. We have access to a sample of $n$ independent and identically distributed (i.i.d.) observations $(Y_i, W_i, X_i)_{i=1,\dots,n}$ and our aim is to identify and estimate the average treatment effect (ATE) $\tau \mathbb{E}[Y(1) - Y(0)]$.

Under the standard assumptions of (a) consistency: $Y = WY(1) + (1 - W)Y(0)$, (b) positivity: there exists $\eta > 0$ such that $1 - \eta \geq \mathbb{P}(W = 1 \,|\, X = x) \geq \eta$ for every $x \in \mathcal{X}$, (c) unconfoundedness: $W \perp\!\!\!\perp \{Y(1), Y(0)\} \,|\, X$, we can nonparametrically identify $\mathbb{E}[Y(w))]$ for instance by the outcome regression/G-formula or the inverse probability weighting (IPW) formula. More precisely, the regression identifiability formula can be obtained as follows:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, |X_i = x]] && \text{(unconfoundedness)} \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i] - \mathbb{E}[Y_i|W_i = 0, X_i]] && \text{(consistency)}
\end{aligned}
$$

This suggests the popular Plug-in G-formula estimator defined as follows.

**Definition 1** (Plug-in G-formula [4])**.** *The plug-in G-formula estimator, denoted by $\hat{\tau}$, is defined as*

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right),$$

*where $\hat{\mu}_w(X)$ is an estimator of $\mu_w(X) = \mathbb{E}\left[Y \mid W = w, X\right]$.*

In this work, we focus on linear outcome models and tackle the challenges of estimating the ATE in a Federated setting using G-formula-based estimators.

## 1.2 Federated learning

When data is collected and stored in distinct centers, it is sometimes impossible to pool them in one place due to privacy or organizational reasons. It is particularly the case when data is sensitive and cannot be shared. For example, hospitals often cannot share their data easily due to medical confidentiality, and often seek to retain control and ownership of their data. On the other hand, gathering enough data to conduct statistically significant observational studies at a single hospital can be challenging. Federated learning [3] aims at addressing these issues by allowing analysts to train a machine learning model across several centers

while keeping data decentralized. Federated learning algorithms typically rely on an orchestrating server and operate across multiple rounds. At each round, (i) the server transmits the current global model to the centers, (ii) each center updates the model using its respective data and sends this model update (not raw data) to the server, and (iii) the server aggregates the updates to form a new global model.

We consider $K$ centers and denote by $Z = \{Z_1, \ldots, Z_K\}$ the pooled dataset, where $Z_k = \{X_k, W_k, Y_k\}$ the data from center $k \in [\![1, K]\!]$ with $n_k$ observations, and $n = \sum_{k=1}^{K} n_k$. The dataset $Z_k$ is only accessible by center $k$. Federated Learning aims at recovering the same results had the researcher had direct access to dataset $Z$.

In the following, we will consider either of the following two assumptions on the sample sizes.

**Assumption 1** (Local Large Sample Size)**.** *For each $w \in \{0, 1\}$ and for each center $k$, we assume $n_k^{(w)} > p$, where $n_k^{(w)}$ is the number of individuals in $\{Z_k \mid W = w\}$ (i.e., the number of treated/control individuals in each center) and $p$ the number of features in $X$.*

**Assumption 2** (Federated Large Sample Size)**.** *For each $w \in \{0, 1\}$, we assume $\sum_{k=1}^{K} n_k^{(w)} > p$.*

In the sequel and unless stated otherwise, we assume that Assumption 1 holds.

### 1.3 Related work

Federated learning has mostly focused on training predictive models [3], and only a handful of recent work considers the problem of federated causal inference. The field of federated causal inference is recent and only few works are available. One can cite [5] who studied the estimation of the ATE in the simplest one-shot federated learning framework (a single round of communication between the server and the centers). In this work, we introduce additional federated estimators and systematically compare the bias and variance of the different strategies under linear models.

## 2 Federated causal inference

### 2.1 Model definition

We consider a random design model, where individuals follow a similar distribution among the $K$ centers. In other words, each center can be seen as a random sample drawn from the same population. We consider the following potential outcomes model with an homogeneous binary treatment effect $\tau \in \mathbb{R}$ with $w \in \{0, 1\}$:

$$Y_i^k(w) = c^{(w)} + X_i^k \beta^{(w)} + \varepsilon_i(w)$$

We denote :

- $Y^k(w) := \{Y_i^k(w)\}_{i=1}^{n_k}$ and $X^k := \{X_i^k\}_{i=1}^{n_k}$ the outcome and covariate vectors for center $k$.
- $Y(w) := \{Y_i(w)\}_{i=1}^{n} = \cup_{k=1}^{K} Y^k(w)$ and $X := \{X_i\}_{i=1}^{n} = \cup_{k=1}^{K} X^k$ the outcome and covariate vectors of the pooled dataset.

and assume:

- (same covariates distribution) $\forall (i,k), X_i^k \sim \mathcal{D}$
- (centered covariates) $\forall k, \mathbb{E}(X) = \mathbb{E}(X^k) = 0$ and $\mathbb{V}(X) = \mathbb{V}(X^k) = \Sigma$
- (regression assumptions) $\forall k, \mathbb{E}(X^{k\top}\varepsilon) = 0, \quad \mathbb{V}(\varepsilon \mid X^k) = \sigma^2, \quad \text{rank}(X^{k\top}X^k) = d$
- (RCT) The treatment variable $W$ is random and does not depend on the covariates: $W \sim \mathcal{B}(p)$.

Finally we define $\tau := \mathbb{E}(Y(1) - Y(0)) = c^{(1)} - c^{(0)} + \mathbb{E}(X)(\beta^{(1)} - \beta^{(0)})$ and $\tau_k := \mathbb{E}(Y^k(1) - Y^k(0)) = c^{(1)} - c^{(0)} + \mathbb{E}(X^k)(\beta^{(1)} - \beta^{(0)})$.

Remark that in this model, we have $\tau_1 = \tau_2 = \cdots = \tau_K = \tau$, which means that the ATE is homogeneous across centers.

Estimating the treatment effect on the pooled data can be done by fitting two ordinary least squares (OLS) regressions, one for the treated and one for the control group. For simplicity, we now - and unless stated otherwise - consider that the first column of $X$ contains the constant vector to serve as intercept, so that The Plug-in G-formula estimator is:

$$\hat{\tau}_{\text{OLS}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{c}^{(1)} + X_i\hat{\beta}_{(1)} - \hat{c}^{(0)} + X_i\hat{\beta}_{(0)}\right)$$

We get the following asymptotic result using the central limit theorem:

$$\sqrt{n}(\hat{\tau}_{\text{OLS}} - \tau) \xrightarrow{d} \mathcal{N}(0, n\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + n\|\beta^{(1)} - \beta^{(0)}\|_{\Sigma}^2) \tag{1}$$

However, the $\hat{\tau}_{\text{OLS}}$ estimator of the ATE cannot be computed in the federated setting since data is decentralized across $K$ centers. Therefore, alternative strategies need to be considered.

## 2.2 Estimation with meta-analysis

A first strategy is to estimate the ATE locally in each center, yielding for each center $k \in [\![1, \ldots, K]\!]$ its associated $\hat{\tau}_k$, and then to aggregate them with weights $\omega_k$ (summing to 1 over the $K$ centers) to get an estimation of the ATE, $\hat{\tau}_{\text{meta}}$, for the whole data, *i.e.*

$$\hat{\tau}_{\text{meta}} = \sum_{k=1}^{K}\omega_k\hat{\tau}_k$$

We call this approach Meta Analysis in reference to the scientific literature that aims at combining the conclusions of multiple studies in order to get a more precise estimate of the quantities of interest. Regarding the aggregation weights, [2] presented two natural choices as we focus on the risk difference: sample size weighting (SW) or weighting the local estimates by the inverse of their variance (IVW).

**Definition 2** (Meta-Analysis with Sample Size Weighting)**.**

$$\hat{\tau}_{\text{meta-SW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k$$

**Property.** *If* $\forall k \in [\![1; K]\!], \mathbb{E}(\hat{\tau}_k) = \tau$*, then* $\hat{\tau}_{\text{meta-SW}}$ *is an unbiased estimator of* $\tau$*.*

**Definition 3** (Meta-Analysis with Inverse Variance Weighting)**.**

$$\hat{\tau}_{\text{meta-IVW}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k)^{-1}.\hat{\tau}_k}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k)^{-1}}$$

**Property.** *If* $\forall k \in [\![1; K]\!], \mathbb{E}(\hat{\tau}_k) = \tau$ *then* $\hat{\tau}_{\text{meta-IVW}}$ *is the minimum-variance unbiased estimator of* $\tau$ *based on aggregation.*

## 2.3 Estimation with one-shot federated learning

A second approach consists in estimating the outcome models parameters in a federated way, using these parameters to estimate the ATE on each center, and ultimately to aggregate these ATE estimates over all centers. In Section 2.3.1, we present an approach to federate the $\hat{\beta}_k$ using a single round of communication ("one-shot" federated learning), following [5]. Then, in Section 2.3.2, we explain how to recover the ATE over all the centers, defining and discussing four estimators of the ATE that rely on the federated parameters.

We define $\hat{\beta}_k^{(w)} := ({X_k^{(w)}}^\top X_k^{(w)})^{-1} {X_k^{(w)}}^\top Y_k^{(w)}$ the local OLS estimator, computed over $Z_k$ and where $X_k^{(w)} = \{X_i^k | W_i^k = w\}_{i=1,\dots,n}$, $Y_k^{(w)} = \{Y_i^k | W_i^k = w\}_{i=1,\dots,n}$ and $w = \{0, 1\}$.

### 2.3.1 One-shot federation of parameters

The first step, which we call the one-shot federation step, is the equivalent of a meta-analysis (local estimation then weighted aggregation) over the $\hat{\beta}_k^{(w)}$. One computes $\hat{\beta}_{\text{fed}}^{(w)} = \sum_{k=1}^{K} \omega_k^{(\beta)} \hat{\beta}_k^{(w)}$ in order to approach the $\hat{\beta}_{\text{pool}}^{(w)}$, with $\omega_k^{(\beta)}$ some chosen federation weights (summing to 1 over the $K$ centers) like sample size weighting and inverse variance weighting.

**Definition 4** (Sample Size Weighting Federation of outcome model parameters)**.**

$$\hat{\beta}_{\text{SW}}^{(w)} = \sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \hat{\beta}_k^{(w)}$$

**Property.** *If* $\forall k \in [\![1;K]\!], \mathbb{E}(\hat{\beta}_k^{(w)}) = \beta^{(w)}$, *then* $\hat{\beta}_{\text{SW}}^{(w)}$ *is an unbiased estimate of* $\beta^{(w)}$

**Definition 5** (Inverse Variance Weighting Federation of outcome model parameters)**.**

$$\hat{\beta}_{\text{IVW}}^{(w)} = \frac{\sum_{k=1}^{K} \left( \mathbb{V}(\hat{\beta}_k^{(w)})^{-1} \hat{\beta}_k^{(w)} \right)}{\sum_{k=1}^{K} \mathbb{V}(\hat{\beta}_k^{(w)})^{-1}}$$

*with* $\mathbb{V}(\hat{\beta}_k^{(w)})^{-1} = \frac{1}{\sigma^2} {X_k^{(w)}}^\top X_k^{(w)}$

**Property** (IVW: Minimum-Variance Estimator)**.** *If* $\forall k \in [\![1;K]\!], \mathbb{E}(\hat{\beta}_k^{(w)}) = \beta^{(w)}$, *then* $\hat{\beta}_{\text{IVW}}^{(w)}$ *is an unbiased estimator of* $\beta^{(w)}$ *with minimum-variance.*

### 2.3.2 Aggregation of the ATE

Once the outcome model parameters have been computed via federation, one can apply the resulting $\hat{\beta}_{\text{fed}}^{(w)} = \{\hat{\beta}_{\text{SW}}^{(w)}$ or $\hat{\beta}_{\text{IVW}}^{(w)}\}$ to each local dataset $Z_k$ to recover $\hat{\tau}_k^{\text{fed}} = \{\hat{\tau}_k^{\text{SW}}$ or $\hat{\tau}_k^{\text{IVW}}\}$, the estimated ATE over center $k$: $\hat{\tau}_k^{\text{fed}} = \frac{1}{n_k} \sum_{i=1}^{n_k} (X_i^k \hat{\beta}_{\text{fed}}^{(1)} - X_i^k \hat{\beta}_{\text{fed}}^{(0)})$.

Finally, we perform an aggregation step of the $\hat{\tau}_k^{\text{fed}}$ to recover $\hat{\tau}_{\text{agg}}^{\text{fed}}$, with a chosen $\omega^{(\tau)}$ aggregation. This yields:

$$\hat{\tau}_{\text{agg}}^{\text{fed}} = \sum_{k=1}^{K} \omega_k^{(\tau)} \hat{\tau}_k^{\text{fed}}$$

Note that we use the superscript "fed" for a federation technique of the outcome model parameters, and the subscript "agg" for the aggregation of the local ATE $\hat{\tau}_k^{\text{fed}}$. We define four estimators of the ATE over the $K$ centers.

**Definition 6** (One-Shot Sample Size Federation - Sample Size Aggregation)**.**

$$\hat{\tau}_{\text{SW}}^{\text{SW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{SW}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_i^k \hat{\beta}_{\text{SW}}^{(1)} - X_i^k \hat{\beta}_{\text{SW}}^{(0)})$$

**Definition 7** (One-Shot Inverse Variance Weighting Federation - Sample Size Aggregation)**.**

$$\hat{\tau}_{\text{SW}}^{\text{IVW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{IVW}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_i^k \hat{\beta}_{\text{IVW}}^{(1)} - X_i^k \hat{\beta}_{\text{IVW}}^{(0)})$$

**Definition 8** (One-Shot Sample Size Federation - Inverse Variance Aggregation)**.**

$$\hat{\tau}_{\text{IVW}}^{\text{SW}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{SW}})^{-1} \cdot \hat{\tau}_k^{\text{SW}}}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{SW}})^{-1}}$$

**Algorithm 1** Federated Averaging

**Input:** $K$ centers, $E$ local epochs, $B$ batch size, $\eta$ the learning rate
**Server executes:**
  Initialize $\beta_0^{(w)}$
**for** each round $t = 0, 1, \dots$ **do**
  **for** each center $k \in [\![1, K]\!]$ **in parallel do**
    $\beta_{t+1}^{k^{(w)}} \leftarrow \text{LocalUpdate}(k, \beta_t^{(w)})$
  **end for**
  $\beta_{t+1}^{(w)} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \beta_{t+1}^{k^{(w)}}$
**end for**
**LocalUpdate**($k$, $\beta_k^{(w)}$):
**for** each local epoch $e = 0, 1, \dots, E-1$ **do**
  $\nabla \ell(\beta_k^{(w)}; \mathcal{D}_k) \leftarrow -\frac{2}{n_k} X^{k^\top} (Y^k - X^k \beta_k^{(w)})$
  $\beta_k^{(w)} \leftarrow \beta_k^{(w)} - \eta \nabla \ell(\beta_k^{(w)}; Z_k)$
**end for**
**return** $\beta_k^{(w)}$

**Definition 9** (One-Shot Inverse Variance Weighting Federation - Inverse Variance Aggregation)**.**

$$\hat{\tau}^{\text{IVW}}_{\text{IVW}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{IVW}})^{-1} \cdot \hat{\tau}_k^{\text{IVW}}}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{IVW}})^{-1}}$$

**Property.** $\hat{\tau}^{\text{IVW}}_{\text{IVW}}$ *is unbiased*

**Property** (Minimum Variance of IVW)**.**

## 2.4 Estimation with Federated Learning

A last strategy consists in estimating the outcome model parameters over the pooled data using the FedAvg algorithm [1], which requires multiple communication rounds. The idea is to estimate $\hat{\beta}^{(w)}_{\text{pool}}$ by solving the underlying OLS problem by gradient descent, which corresponds to minimizing the Mean Squared Error loss function $\ell(\beta^{(w)}; Z_1^{(w)}, \dots, Z_K^{(w)}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \beta)^2$. The algorithm, shown in Algorithm 1, alternates between local gradient steps in each center and aggregation at the server, to eventually produce an estimate $\hat{\beta}^{(w)}_{\text{GD}}$ ("GD" stands for Gradient Descent).

Under an appropriate choice of hyper-parameters, the FedAvg algorithm converges [3] and we have $\hat{\beta}^{(w)}_{\text{GD}} = \hat{\beta}^{(w)}_{\text{pool}}$. Once $\hat{\beta}^{(w)}_{\text{GD}}$ is obtained, we compute $\hat{\tau}_k^{\text{GD}} := \frac{1}{n_k} \sum_{i=1}^{n_k} \left( X_i^k \hat{\beta}^{(1)}_{\text{GD}} - X_i^k \hat{\beta}^{(0)}_{\text{GD}} \right)$, and then calculate the estimates of the ATE over the $K$ centers $\hat{\tau}^{\text{GD}}_{\text{agg}} = \sum_{k=1}^{K} \omega_k^{(\tau)} \hat{\tau}_k^{\text{GD}}$ with a chosen aggregation method as in Section 2.3.

**Definition 10** (Gradient Descent Federation - Sample Size Aggregation)**.**

$$\hat{\tau}^{\text{GD}}_{\text{SW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}^{\text{GD}}_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left( X_i^k \hat{\beta}^{(1)}_{\text{GD}} - X_i^k \hat{\beta}^{(0)}_{\text{GD}} \right)$$

**Definition 11** (Gradient Descent Federation - Inverse Variance Aggregation)**.**

$$\hat{\tau}^{\text{GD}}_{\text{IVW}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}^{\text{GD}}_k)^{-1} \hat{\tau}^{\text{GD}}_k}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}^{\text{GD}}_k)^{-1}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}^{\text{GD}}_k)^{-1} \frac{1}{n_k} \sum_{i=1}^{n_k} \left( X_i^k \hat{\beta}^{(1)}_{\text{GD}} - X_i^k \hat{\beta}^{(0)}_{\text{GD}} \right)}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}^{\text{GD}}_k)^{-1}}$$

**Property.** *If* $\mathbb{E}(\hat{\beta}^{(w)}_{\text{GD}}) = \beta^{(w)}$, $\hat{\tau}^{\text{GD}}_{\text{SW}}$ *and* $\hat{\tau}^{\text{GD}}_{\text{SW}}$ *are unbiased.*

## 2.5 Properties, Bias and Variance of the Federated Estimators

Under the random design, one can derive asymptotic properties of the estimators, summarized in this table.

| Estimator | Notation | Definition | $\mathbb{E}$ | $\mathbb{V}^{\infty}$ |
|---|---|---|---|---|
| Pool | $\hat{\beta}^{(w)}_{\text{pool}}$ | $\left( X^{(w)\top} X^{(w)} \right)^{-1} X^{(w)\top} y^{(w)}$ | $\beta^{(w)}$ | $\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$ |
| Local | $\hat{\beta}^{(w)}_k$ | $\left( X_k^{(w)\top} X_k^{(w)} \right)^{-1} X_k^{(w)\top} y_k^{(w)}$ | $\beta^{(w)}$ | $\frac{\sigma^2}{n_k^{(w)}} \Sigma^{-1}$ |
| Federated | $\hat{\beta}^{(w)}_{\text{GD}}$ | obtained by gradient descent | $\beta^{(w)}$ | $\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$ |
| Sample Size Weighted | $\hat{\beta}^{(w)}_{\text{SS}}$ | $\sum_{k=1}^{K} \frac{n_k}{n} \hat{\beta}^{(w)}_k$ | $\beta^{(w)}$ | $\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$ |
| Inverse Variance Weighted | $\hat{\beta}^{(w)}_{\text{IVW}}$ | $\frac{\sum_{k=1}^{K} \left( \mathbb{V}^{\infty}(\hat{\beta}^{(w)}_k)^{-1} \hat{\beta}^{(w)}_k \right)}{\sum_{k=1}^{K} \mathbb{V}^{\infty}(\hat{\beta}^{(w)}_k)^{-1}}$ | $\beta^{(w)}$ | $\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$ |

Table 1: Estimators of the outcome model parameters

**Property 1.** $\forall k \in [\![1; K]\!]$, $\frac{n_k}{n} = \frac{\mathbb{V}^{\infty}(\hat{\tau}_k \text{fed})^{-1}}{\sum_{k=1}^{K} \mathbb{V}^{\infty}(\hat{\tau}_k \text{fed})^{-1}}$, *which yields* $\hat{\tau}^{\text{fed}}_{\text{SW}} = \hat{\tau}^{\text{fed}}_{\text{IVW}}$

Then, using that $\sqrt{n}(\hat{\tau}_{\text{OLS}} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \left( \frac{1}{n^1} + \frac{1}{n^0} \right) + \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma} \right)$, we get the following asymptotic results for the ATE estimators:

| Estimator | Notation | Definition | $\mathbb{E}$ | $\mathbb{V}^\infty$ |
|---|---|---|---|---|
| Pool | $\hat{\tau}_{\text{pool}}$ | $\frac{1}{n}\sum_{i=1}^n X_i\left(\hat{\beta}_{\text{pool}}^{(1)^\top} - \hat{\beta}_{\text{pool}}^{(0)^\top}\right)$ | $\tau$ | $\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| Local | $\hat{\tau}_k$ | $\frac{1}{n}\sum_{i=1}^n X_i\left(\hat{\beta}_k^{(1)^\top} - \hat{\beta}_k^{(0)^\top}\right)$ | $\tau$ | $\sigma^2\left(\frac{1}{n_k^{(1)}} + \frac{1}{n_k^{(0)}}\right) + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| fed-Local | $\hat{\tau}_k^{\text{fed}}$ | $\frac{1}{n}\sum_{i=1}^n X_i\left(\hat{\beta}_{\text{fed}}^{(1)^\top} - \hat{\beta}_{\text{fed}}^{(0)^\top}\right)$ | $\tau$ | $\frac{\sigma^2}{n_k}\left(\frac{n}{n^{(1)}} + \frac{n}{n^{(0)}}\right) + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| meta-SW | $\hat{\tau}_{\text{SW}}$ | $\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k$ | $\tau$ | $\sum_{k=1}^K\left(\frac{n_k}{n^2}\left(\left(\frac{\sigma^2 n_k}{n_k^{(1)}} + \frac{\sigma^2 n_k}{n_k^{(0)}}\right) + \|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2\right)\right)$ |
| meta-IVW | $\hat{\tau}_{\text{IVW}}$ | $\frac{\sum_{k=1}^K\left(\mathbb{V}^\infty(\hat{\tau}_k)^{-1}\hat{\tau}_k\right)}{\sum_{k=1}^K \mathbb{V}^\infty(\hat{\tau}_k)^{-1}}$ | $\tau$ | $\left(\sum_{k=1}^K \mathbb{V}^\infty(\hat{\tau}_k)^{-1}\right)^{-1}$ |
| fed-SW-agg | $\hat{\tau}_{\text{SW}}^{\text{fed}}$ | $\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k^{\text{fed}}$ | $\tau$ | $\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| fed-IVW-agg | $\hat{\tau}_{\text{IVW}}^{\text{fed}}$ | $\frac{\sum_{k=1}^K\left(\mathbb{V}^\infty(\hat{\tau}_k^{\text{fed}})^{-1}\hat{\tau}_k^{\text{fed}}\right)}{\sum_{k=1}^K \mathbb{V}^\infty(\hat{\tau}_k^{\text{fed}})^{-1}}$ | $\tau$ | $\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |

Table 2: Estimators of the ATE in random design, RCT

One can rewrite the table above for the special RCT case of a completely balanced design, where $n_k^{(1)} = n_k^{(0)} = n_k/2$ for all $k \in [\![1, K]\!]$:

| Estimator | Notation | Definition | $\mathbb{E}$ | $\mathbb{V}^\infty$ |
|---|---|---|---|---|
| Pool | $\hat{\tau}_{\text{pool}}$ | $\frac{1}{n}\sum_{i=1}^n\left(X_i\hat{\beta}_1^{\text{pool}\top} - X_i\hat{\beta}_{(0)}^{\text{pool}\top}\right)$ | $\tau$ | $\frac{4\sigma^2}{n} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| Local | $\hat{\tau}_k$ | $\frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^k\hat{\beta}_k^{(1)\top} - X_i^k\hat{\beta}_k^{(0)\top}\right)$ | $\tau$ | $\frac{4\sigma^2}{n_k} + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| fed-Local | $\hat{\tau}_k^{\text{fed}}$ | $\frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^k\hat{\beta}_1^{\text{fed}\top} - X_i^k\hat{\beta}_{(0)}^{\text{fed}\top}\right)$ | $\tau$ | $\frac{4\sigma^2}{n_k} + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| meta-SW | $\hat{\tau}_{\text{SW}}$ | $\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k$ | $\tau$ | $\frac{4\sigma^2}{n} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| meta-IVW | $\hat{\tau}_{\text{IVW}}$ | $\frac{\sum_{k=1}^K\left(\mathbb{V}(\hat{\tau}_k)^{-1}\hat{\tau}_k\right)}{\sum_{k=1}^K \mathbb{V}(\hat{\tau}_k)^{-1}}$ | $\tau$ | $\frac{4\sigma^2}{n} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| fed-SW-agg | $\hat{\tau}_{\text{SW}}^{\text{fed}}$ | $\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k^{\text{fed}}$ | $\tau$ | $\frac{4\sigma^2}{n} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |
| fed-IVW-agg | $\hat{\tau}_{\text{IVW}}^{\text{fed}}$ | $\frac{\sum_{k=1}^K\left(\mathbb{V}(\hat{\tau}_k^{\text{fed}})^{-1}\hat{\tau}_k^{\text{fed}}\right)}{\sum_{k=1}^K \mathbb{V}(\hat{\tau}_k^{\text{fed}})^{-1}}$ | $\tau$ | $\frac{4\sigma^2}{n} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$ |

Table 3: Estimators of the ATE in fixed design

# 3 Conclusion

In this paper, we have defined and studied several estimators for the ATE in the context of federated causal inference. We have provided a comprehensive comparison of the variance of these estimators, and we have shown that in the simple case of homogeneous settings with no centre-effect, the meta-analysis estimators are asympotically as efficient as the federated estimators. Simulation results that are not presented in this paper also back up this finding.

The natural next steps of this work would be to explore the context of heterogeneous settings, where populations or treatment assignment rules are different from centers to centers. Heterogeneous treatment effects as well as nonlinear models are interesting horizons

and other estimators of the ATE like the Augmented Inverse Propensity Score (AIPW) can also be studied. Covariate mismatch between centers and high dimensionality are important topics to consider, especially in the context of hospital data. Finally, the development of federated causal inference methods that can be applied to real-world data is an important direction for future research.

# References


[1] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[2] John E Hunter and Frank L Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

[3] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, and Mehdi Bennis et al. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[4] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[5] Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*, 2021.

# Multi-omique

# COMPARATIVE ANALYSIS OF SUPERVISED INTEGRATIVE METHODS FOR MULTI-OMICS DATA

**Alexei Novoloaca,[1,*] Camilo Broc,[1] Laurent Beloeil,[1] Wen-Han Yu[2] and Jérémie Becker[1]**

**[1], BIOASTER Research Institute, 40 avenue Tony Garnier, 69007, Lyon, France**

**[2], Bill & Melinda Gates Medical Research Institute, Cambridge, Massachusetts, USA**

***Corresponding author. alexei.novoloaca@bioaster.org**

**Camilo Broc: Camilo.Broc@bioaster.org**

**Laurent Beloeil: Laurent.Beloeil@bioaster.org**

**Wen-Han Yu: wenhan.yu.dr@gmail.com**

**Jérémie Becker: Jeremie.Becker@bioaster.org**

**Résumé**

Les récents progrès dans les technologies de séquençage, de spectrométrie de masse et de cytométrie ont permis de collecter plusieurs types de données omiques à partir d'un seul échantillon. Ces vastes ensembles de données ont conduit à un consensus croissant selon lequel une approche holistique est nécessaire pour identifier de nouveaux biomarqueurs candidats et dévoiler les mécanismes sous-jacents à l'étiologie des maladies, clé de la médecine de précision. Bien que de nombreuses revues et évaluations aient été menées sur les approches non supervisées (Bersanelli et al., 2016), leurs homologues supervisés ont reçu moins d'attention dans la littérature et aucun standard de référence n'a encore émergé (Krassowski et al., 2020).

Dans ce travail, nous présentons une comparaison approfondie d'une sélection de cinq méthodes, représentatives des principales familles d'approches intégratives (factorisation de matrice, méthodes à noyaux multiples, apprentissage ensembliste et méthodes basées sur les graphes). Comme contrôle non-intégratif, une forêt d'arbre de décision a été exécutée sur des ensembles de données concaténés et séparés. Les méthodes ont été évaluées à la fois sur des données simulées et réelles, ces dernieres étant soigneusement sélectionnées pour couvrir différentes applications médicales (maladies infectieuses, oncologie et vaccins) et différentes types d'omiques. Un ensemble de quinze scénarios de simulation a été conçu à partir des données réelles pour explorer un espace de paramétres vaste et réaliste (par exemple, taille de l'échantillon, dimensionnalité, déséquilibre de classes, taille de l'effet).

La comparaison entre les approches intégratives et non intégratives a montré que bien que des performances comparables aient été trouvées sur les simulations, les méthodes basées sur des variables latentes surpassaient généralement les méthodes non-intégratives sur les données réelles. Les forces et les limites de ces méthodes seront discutées en détail ainsi que des lignes directrices pour les futures applications.

**Mots-clés : analyse comparative ; intégration de données ; données multi-omiques ; modèles de prédiction ; analyse supervisée**

**Abstract**

Recent advances in sequencing, mass spectrometry and cytometry technologies have enabled researchers to collect multiple 'omics data types from a single sample. These large datasets have led to a growing consensus that a holistic approach is needed to identify new candidate biomarkers and unveil mechanisms underlying disease aetiology, a key to precision medicine. While many reviews and benchmarks have been conducted on unsupervised approaches (Bersanelli et al. 2016 [1]), their supervised counterparts have received less attention in the literature and no gold standard has emerged yet (Krassowski et al. 2020 [2]).

In this work, we present a thorough comparison of a selection of five methods, representative of the main families of integrative approaches (matrix factorization, multiple kernel methods, ensemble learning and graph-based methods). As non-integrative control, random forest was performed on concatenated and separated data types. Methods were evaluated both on simulated and real-world datasets, the latter being carefully selected to cover different medical applications (infectious diseases, oncology and vaccine) and data modalities. A set of fifteen simulation scenarios were designed from the real-world datasets to explore a large and realistic parameter space (*e.g.* sample size, dimensionality, class imbalance, effect size).

The comparison between integrative and non-integrative approaches showed that although comparable performances were found on simulations, latent variable models generally outperformed non-integrative methods on experimental data. The strengths and limitations of these methods will be discussed in detail as well as guidelines for future applications.

**Key words: benchmark; data integration; multi-omics data; prediction models; supervised analysis**

## Benchmark workflow

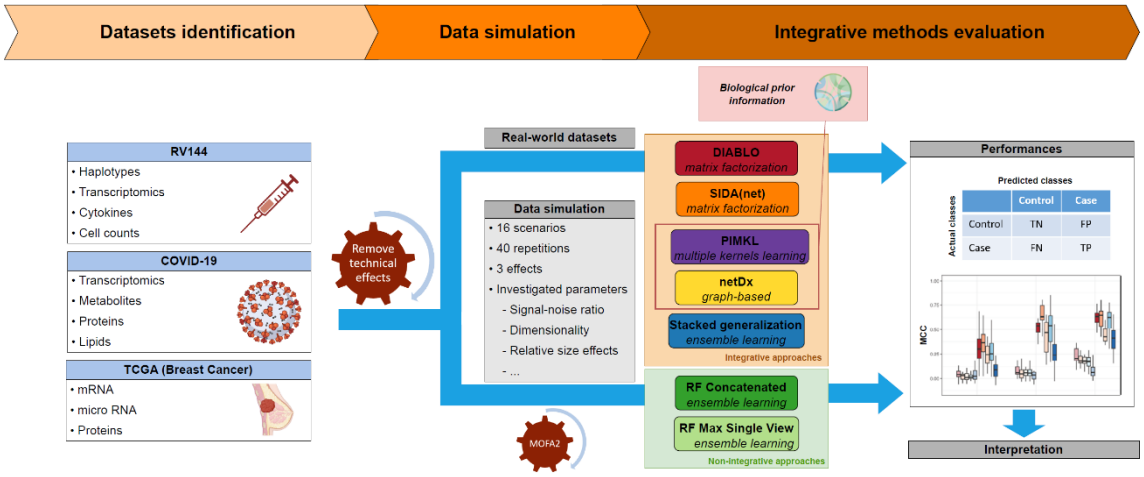



*Figure 1 Three multi-omics datasets, covering distinct medical applications, were selected. A reference simulation scenario was designed using signal-to-noise ratio (SNR) and sparsity levels estimated from real-world datasets. 14 alternative scenarios were also generated by modifying class imbalance, SNR, dimensionality, relative importance of effects, etc. A selection of five integrative approaches, representative of existing methods, were evaluated on both real-world and simulated data based using MCC.*

### Context

The continuous progress made in omics technologies have reshaped our understanding of human biology. While single omics analyses have produced valuable insights, most common human diseases associated with high mortality (e.g. type 2 diabetes, cardiovascular disease) still lack effective therapeutic strategies. Moreover, diverse large-scale cancer projects, such as TCGA [3], ICGC [4], COSMIC [5], consistently demonstrated the power of data integration in patient stratification.

With a growing interest in extracting multi-omics features associated with health-related outcomes, an in-depth understanding of current supervised integrative approaches is much needed. In this context, a wide variety of integrative approaches have been introduced to address one or some of the following goals: (i) patient stratification, (ii) prediction of clinical outcome and (iii) identification of molecular mechanisms acting across molecular layers.

## Methods

The current literature commonly distinguishes six main families of methods:

- matrix factorization
- multiple kernel learning
- network-based methods
- bayesian
- ensemble learning
- deep learning

We selected five methods (see Table 1) representative of the main families of integrative approaches (matrix factorization, multiple kernel methods, ensemble learning and graph-based methods) that we evaluated on a set of 15 simulations (see Table 2) exploring a large and realistic parameter space (e.g. sample size, dimensionality, confounding effects, effect size) and on real-world datasets. The datasets choosen were carefully selected to cover different medical applications (infectious diseases, oncology and vaccine).

As non-integrative control, two alternatives based on Random Forest (RF) were also included in this benchmark to evaluate the added value of data integration: RF_Concat concatenates omics layers sample-wise and evaluate the overall performance, while RF_Max_Single_View consists in evaluating RF on each modality and keeping only the highest classification performance.

| | Name | Underlying approach | Prior information | Implementation |
|---|---|---|---|---|
| Integrative methods | DIABLO | Sparse generalized CCA | No | R package *mixOmics* |
| | SIDANet | Combination of LDA and CCA | Yes | R package *SIDA* |
| | PIMKL | Multiple kernel learning | Yes | Python script |
| | netDx | Integrated patient Similarity network | Yes | R package *netDx* |
| | Stacked generalization | Ensemble of weak learners | No | R package *SuperLearner* |
| Non-integrative methods | RF_Concat<br>RF_Max_Single_View | Random Forest on concatenated data<br>separated data | No | R package *randomForest* |

*Table 1 Summary of the methods selected in the benchmark.*

## Classification performance criterion

The Matthews Correlation Coefficient (MCC) is a popular metric used to evaluate the performance of binary classifiers. In a recent study, Chicco *et al.* [6] advocated for its use over accuracy and F1-score due to its robustness in imbalanced setting and invariance for class swapping. The MCC is defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

where TP and TN are the number of true positives and true negatives; FP and FN the number of false positives and false negatives. Like Pearson's correlation coefficient, MCC ranges between -1 and 1. ±1 reflects perfect (mis)classification, while 0 indicates random classification. To ensure an unbiased comparison, all methods were evaluated in 5-fold cross-validation.

## Simulation scenarios

| Scenario | Number of samples (cases, controls) | Number of features per omic | Main factor(s) | Fraction of signal features per omic | Overlap across factors |
|---|---|---|---|---|---|
| **Reference** | **80 (40, 40)** | **1000, 240, 60** | **All equal** | **0.1, 0.1, 0.1** | **0.5** |
| n/5 | 16 (8, 8) | | | | |
| px5 | | 5000, 1200, 300 | | | |
| CaseControl_1:7 | 80 (10, 70) | | | | |
| High_Main_MO | | | High main MO | | |
| High_Conf_MO_Overlap | | | High confounding MO | | 0.95 |
| Main_MO_2Smallest_Omics | | | Main MO kept in the | | |
| Main_MO_1Largest_Omic | | | smallest/largest omic(s) | | |
| High_Fraction_Signal_Feat | | | | 0.3, 0.3, 0.3 | |
| nx5 | 400 (200, 200) | | | | |
| p/5 | | 200, 48, 12 | | | |
| High_Conf_SO_Overlap | | | High confounding SO | | 0.95 |
| Main_MO_1Smallest_Omic | | | Main MO kept in the | | |
| Main_MO_2Largest_Omics | | | smallest/largest omic(s) | | |
| Noise | | | | 0, 0, 0 | |

*Table 2 15 scenarios were generated from real-world datasets. The reference scenario is defined by 2 classes of 40 samples each, 3 omics with p = (1000, 240, 60) variables and 3 factors (a main multi-omics, Main MO and two confounding factors acting at the single- and multi-omics levels, Conf SO, Conf MO). For the other scenarios, only the deviations from the reference are indicated.*

## Results on simulated data

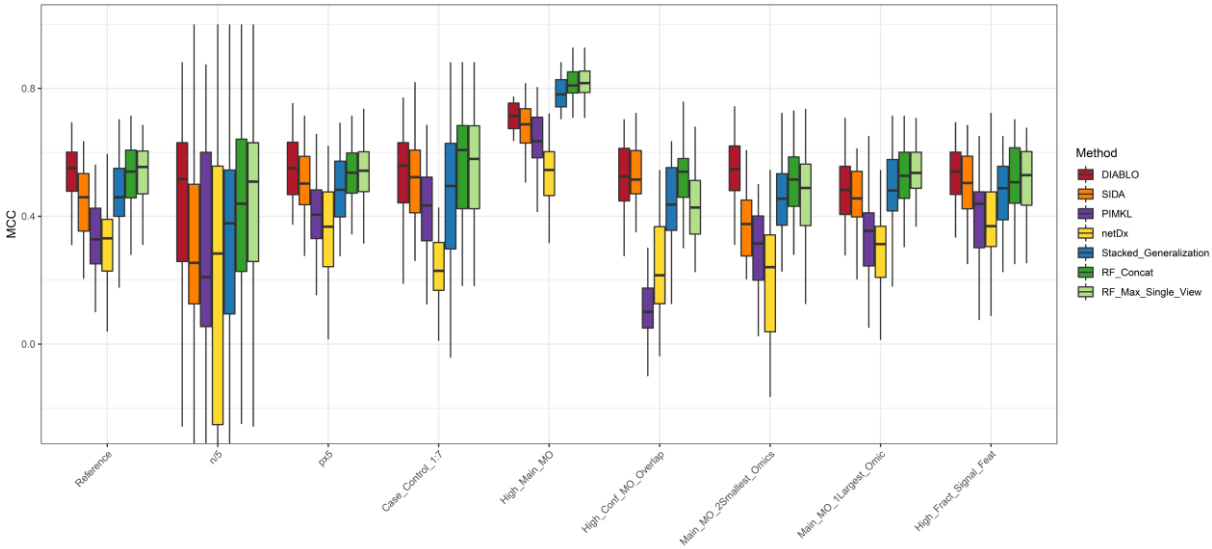



*Figure 2 Method comparison on simulated data. Integrative approaches were evaluated on main 9 simulation scenarios. Two non-integrative methods (RF_Concat, RF_Max_Single_View) were also included to quantify the added-value of data integration. For each scenario, 40 repetitions were generated, on which, MCC was computed in 5-fold cross-validation.*

## Results on real-world data

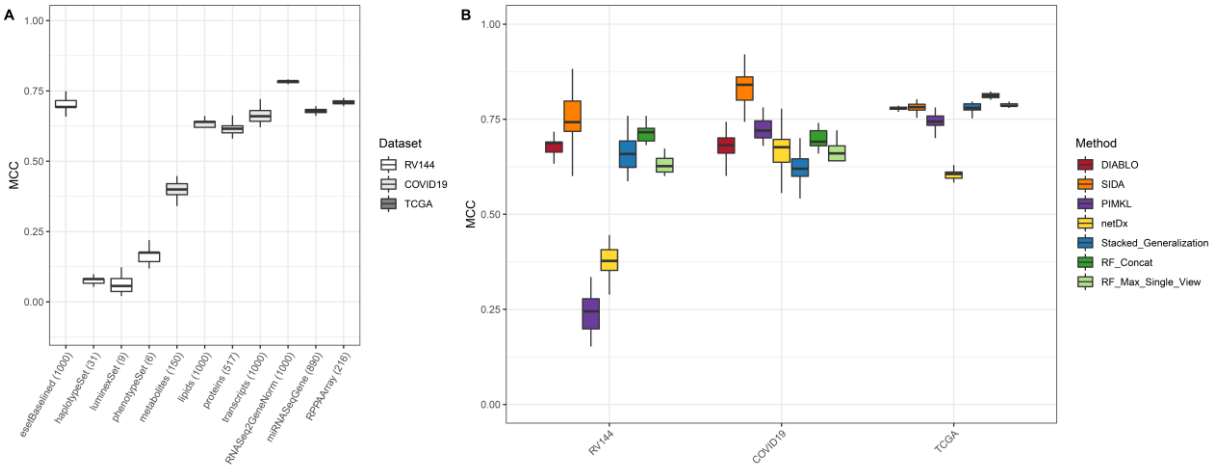



*Figure 3 Method comparison on 3 real-world datasets. Prediction performance (A) on individual omic using Random Forest or (B) integrative methods. MCC was computed on 40 repetitions of 5-fold cross-validation.*

## Key points

- Supervised integrative methods have received little attention in the literature so far. In this work, five supervised methods spanning major families of integrative approaches are thoroughly evaluated.
- Non-integrative approaches (Random Forest based) were further included to elucidate the conditions in which data integration provides a clear advantage.
- In many simulation scenarios, Random Forest and latent variable models lead to comparable performances.
- When the main multi-omics effect is present in a subset of views, integrative approaches demonstrate their superiority. Conversely, when the multi-omics effect is strong, Random Forest outperforms its integrative counterparts.
- Among integrative approaches, latent variable models lead to the best performance on simulated (DIABLO) and experimental (SIDA) data.

## References


1. Matteo Bersanelli et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics, 17(S2):S15, December 2016.
2. Michal Krassowski et al. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Frontiers in Genetics, 11:610798, December 2020.
3. Katarzyna Tomczak, Patrycja Czerwi´nska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Wsp´olczesna Onkologia, 2015(1):68–77, 2015. Publisher: Termedia.
4. Thomas J. Hudson (Chairperson), Warwick Anderson, Axel Aretz, et al. International network of cancer genome projects. Nature, 464(7291):993–998, April 2010. Number: 7291 Publisher: Nature Publishing Group.
5. John G. Tate, Sally Bamford, Harry C. Jubb, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research, 47(D1):D941–D947, January 2019.
6. Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1):6, January 2020.

# Utilisation de la NMF supervisée intégrative pour l'étude d'altérations de la peau

Aurélie Mercadié[1,2], Éléonore Gravier[2], Gwendal Josse[2], Nathalie Vialaneix[1] & Céline Brouard[1]

[1] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France,* {*aurelie.mercadie, nathalie.vialaneix, celine.brouard*}*@inrae.fr*
[2] *Laboratoires Pierre Fabre, F-31300, Toulouse, France,* {*eleonore.gravier, gwendal.josse*}*@pierre-fabre.com*

**Résumé.** Cette communication est motivée par un problème fréquent en recherche clinique : des patients, stratifiés en groupes d'intérêt (typiquement sains / malades ou contrôles / traités) sont caractérisés par des mesures biologiques correspondant à plusieurs caractéristiques différentes (métabolomiques, protéomiques, etc). L'objectif est alors de découvrir des signatures moléculaires multi-omiques caractérisant les groupes.

Ici, nous proposons une approche de Factorisation Matricielle Non-négative (NMF) que nous étendons à ce cadre. De manière plus précise, notre proposition se fonde sur une variante du problème d'optimisation FR-lda [10], qui, par l'introduction d'un terme supervisé, permet de prendre en compte la structuration des individus en groupes d'intérêt. Notre proposition étend cette méthode à un cadre multi-tableaux, en assurant l'intégration des informations par le biais d'une composante de pondération commune à tous les tableaux. Nous proposons deux approches pour résoudre le problème d'optimisation induit, l'une classique, par approche multiplicative, et l'autre, nouvelle, par approche proximale et qui permet d'obtenir une parcimonie exacte dans les signatures moléculaires.

Nous illustrerons l'utilisation de cette extension sur des données de protéomique et de transcriptomique issues d'échantillons de peau prélevés sur une zone non-lésionnelle de sujets sains et de sujets atteints de Dermatite Atopique (DA). Nous montrerons comment celle-ci nous a permis d'identifier une signature multi-omique de la DA.

**Mots-clés.** intégration multi-omiques, apprentissage supervisé, NMF, optimisation proximale.

**Abstract.** This communication is motivated by a frequent problem in clinical research: patients, stratified into groups of interest (typically healthy/sick or control/treated patients) are described by biological measurements corresponding to different omics (metabolomics, proteomics, etc.). The aim is then to discover molecular signatures characterizing the groups.

Here, we propose a Non-negative Matrix Factorization (NMF) approach that we extend to this framework. More specifically, our proposal is based on the FR-lda variant of NMF [10]. This method introduces a supervised term, aiming at explaining the two groups of individuals. Our proposal extends this method to a multi-table framework, by integrating information through a weighting matrix common to all omics. We also propose two approaches to solve the induced optimization problem, the classical multiplicative approach (MU) and a novel proximal approach that achieves exact sparsity in molecular signatures.

The use of this extension is illustrated on proteomic and transcriptomic data from skin samples taken in a non-lesional area of subjects with and without Atopic Dermatitis (DA). First results show that our NMF variant identifies a multi-omic signature of the disease.

**Keywords.** multi-omics integration, supervised learning, NMF, proximal optimization.

# 1 Introduction

Cette communication est motivée par un problème fréquent en recherche clinique : des patients, stratifiés en groupes d'intérêt (typiquement sains / malades ou contrôles / traités) sont caractérisés par des mesures biologiques correspondant à plusieurs caractéristiques différentes (métabolomiques, protéomiques, etc). L'objectif est alors de découvrir des signatures moléculaires multi-omiques caractérisant les groupes. En particulier, les Laboratoires Pierre Fabre sont engagés dans de multiples projets de ce type dans lesquels des données omiques multiples ont été acquises dans le but de mieux comprendre des altérations de la peau. Ici, nous nous focalisons en particulier sur un projet lié à la dermatite atopique (DA) qui est une maladie inflammatoire commune, principalement caractérisée par une fonction barrière de la peau dysfonctionnelle. Plusieurs études multi-omiques portant sur des échantillons de peau humaine ont pu caractériser la DA sur le plan protéomique ou transcriptomique [7, 13, 3], or les analyses séparées de ces différentes omiques ne permettent pas de comprendre les relations gènes-protéines potentiellement instrumentales dans le développement de cette maladie.

Or, si les approches permettant l'intégration de données multiples, en particulier de données omiques multiples, se sont développées de manière importante ces dernières années (voir notamment [12, 11, 4] pour des revues sur ce sujet), un faible nombre sont destinées à une analyse exploratoire tenant compte de cette structure. Ici, nous abordons donc la question de l'intégration multi-omiques sous un angle mixte, celui de l'analyse exploratoire d'omiques multiples dans laquelle une information complémentaire caractérisant ces individus (attribut clinique ou expérimental par exemple) est d'intérêt pour la compréhension du phénomène biologique.

Dans cette communication, nous présentons une extension de la Factorisation Matricielle Non-négative (NMF) [8] et plus particulièrement de sa version supervisée [9] pour extraire un profil multi-omique de la DA. En effet, cette méthode de réduction de dimension offre un cadre bien adapté au problème d'intégration de données omiques pour des individus structurés en groupe. Également, elle est spécifiquement conçue pour l'analyse de données à valeurs positives, ce qui est le cadre naturel de nombreuses données omiques (données de comptages comme le RNA-seq ou les données métagénomiques, données compositionnelles comme en métabolomique ou protéomique, etc). En particulier, son interprétation est elle-même facilitée par la contrainte de positivité de la solution, la décomposition retenue s'expliquant aisément en termes de profils types et d'appartenance à ces profils de chacun des individus.

Dans la suite, nous introduisons le cadre général de la NMF, certaines variantes ainsi que l'approche que nous proposons, une version intégrative et supervisée de la NMF, dans

la section 2. Enfin, dans la section 3, nous présentons les premiers résultats des expériences obtenues sur données simulées et sur les données du projet étudiant la dermatite atopique.

# 2 NMF supervisée intégrative

## 2.1 La NMF et ses variantes

Soit $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ une matrice de données à entrées positives de grande dimension (typiquement $n \ll p$). Le but de la NMF est de fournir une approximation de faible rang de $\mathbf{X}$ telle que :

$$\mathbf{X} \simeq \mathbf{W} \times \mathbf{H},$$

avec $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ et $\mathbf{H} \in \mathbb{R}_+^{K \times p}$, deux matrices à entrées positives respectivement appelées matrice des poids et matrice des signatures (ou composantes latentes). $K$ est le nombre de signatures (ou le rang de la décomposition) et est choisi par l'utilisateur.

[8] décrivent deux variantes de la NMF qui se fondent sur deux fonctions de coût distinctes dont le but est de minimiser l'erreur de l'approximation : la divergence de Kullback-Leibler et la norme de Fröbenius. Le choix de la fonction de coût dépend de la distribution statistique que l'on attribue au terme d'erreur et pour des distributions gaussiennes, la norme de Fröbenius est généralement indiquée :

$$\underset{\mathbf{W},\mathbf{H} \geq 0}{\arg\min} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 .$$

Ainsi, la NMF est initialement une méthode non supervisée, conçue pour les analyses exploratoires. Toutefois, [10] ont développé plusieurs variantes de la NMF adaptées à la classification d'images MALDI[1], dont la NMF « FR-lda ». Cette variante intègre un terme supervisé qui assure que la décomposition obtenue (en particulier les signatures) est prédictive d'une structuration des individus en deux groupes, $\mathbf{y} \in \{0, 1\}^n$ :

$$\underset{\mathbf{W},\mathbf{H},\boldsymbol{\beta} \geq 0}{\arg\min} \underbrace{\frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \frac{\mu}{2} \|\mathbf{W}\|_F^2 + \lambda \|\mathbf{H}\|_1 + \frac{\nu}{2} \|\mathbf{H}\|_F^2 + \frac{\gamma}{2} \|\mathbf{y} - \mathbf{X}\mathbf{H}^\top \boldsymbol{\beta}\|_2^2}_{= \mathcal{F}_0(\mathbf{W},\mathbf{H},\boldsymbol{\beta})}$$

avec :

- $\mathbf{W} \in \mathbb{R}_+^{n \times K}$, la matrice des poids (contribution des individus aux composantes latentes) ;
- $\mathbf{H} \in \mathbb{R}_+^{K \times p}$, les signatures (composantes latentes) ;
- $\boldsymbol{\beta} \in \mathbb{R}_+^K$, les coefficients de régression ;
- $\mu$, $\lambda$, $\nu$, $\gamma > 0$, les paramètres de régularisation, fixés.

[1] Matrix Assisted Laser Desorption/Ionization

Outre le terme classique de perte de l'approximation et le terme supervisé correspondant à un critère de moindres carrés, $\|\mathbf{y} - \mathbf{X}\mathbf{H}^\top \boldsymbol{\beta}\|_2^2$, les termes de perte $\ell_2$ assurent la régularisation de la solution (ainsi que la définition d'une perte trivialement non identifiable) et le terme de perte $\ell_1$ assure la parcimonie des signatures obtenues.

## 2.2 Extension de la NMF FR-lda pour l'intégration de données

Dans le cadre multi-omique décrit dans l'introduction, nous considérons maintenant $\mathbf{X}^{(j)} \in \mathbb{R}_+^{n \times p_j}$ ($j \in \{1, 2\}$), deux matrices à entrées positives contenant les mesures de deux types d'omiques sur les mêmes $n$ individus ($p_j$ étant le nombre de variables mesurées dans chaque omique). On notera également $\mathbf{y} \in \{0, 1\}^n$ le vecteur d'appartenance des individus à deux groupes d'intérêt biologique.

Notre proposition consiste à étendre l'approche de [6] pour l'intégration de données en minimisant le critère :

$$\begin{aligned} \underset{\mathbf{W}, \mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}}{\arg\min} \frac{1}{2} \left( \sum_{j=1}^{2} \|\mathbf{X}^{(j)} - \mathbf{W}\mathbf{H}^{(j)}\|_F^2 \right) + \frac{\gamma}{2} \left( \sum_{j=1}^{2} \|\mathbf{y} - \mathbf{X}^{(j)}\mathbf{H}^{(j)\top}\boldsymbol{\beta}^{(j)}\|_2^2 \right) \\ + \sum_{j=1}^{2} \lambda \|\mathbf{H}^{(j)}\|_1 + \frac{\mu}{2} \|\mathbf{W}\|_F^2 \end{aligned} \tag{1}$$

avec :

- $\mathbf{W} \in \mathbb{R}_+^{n \times K}$, la matrice des poids, commune aux deux composantes latentes ;
- $\forall j \in \{1, 2\}$, $\mathbf{H}^{(j)} \in \mathbb{R}_+^{K \times p_j}$, les signatures, spécifiques de chaque omique ;
- $\forall j \in \{1, 2\}$, $\boldsymbol{\beta}^{(j)} \in \mathbb{R}_+^K$, les coefficients de régression ;
- $\gamma$, $\lambda$, $\mu > 0$, les paramètres de régularisation, fixés.

Les problèmes d'optimisation qui apparaissent dans la NMF sont des problèmes non-convexes et non-linéaires [6] car la fonction de perte n'est pas convexe en $\mathbf{W}$, $\mathbf{H}^{(1)}$, $\mathbf{H}^{(2)}$, $\boldsymbol{\beta}^{(1)}$ et $\boldsymbol{\beta}^{(2)}$ simultanément. Toutefois, la marginalisation en chacune de ces variables conduit à des problèmes d'optimisation convexes, deux de ces problèmes incluant une contrainte non lisse (la pénalité $\ell_1$). Les problèmes d'optimisation de type NMF sont donc généralement résolus par des approches itératives résolvant successivement chacun des problèmes marginaux en utilisant des méthodes de Majoration-Minimisation (MM). En outre, dans le cas particulier de la NMF, ce principe permet d'obtenir des étapes de mises à jour multiplicatives (MU) qui assurent la positivité des matrices obtenues à chaque itération. Toutefois, ces méthodes ne permettent qu'une parcimonie approchée. Alternativement, nous avons proposé une méthode permettant la parcimonie exacte des signatures en remplaçant l'étape MU par une étape d'optimisation proximale. De manière générale, le problème de l'équation (1) est résolu par l'algorithme 1.

**Algorithme 1** Vue d'ensemble de l'algorithme utilisé pour la résolution de l'équation (1)

1: Initialiser les matrices $\mathbf{W}^{(0)}$, $\mathbf{H}^{(j,0)}$ et vecteurs $\boldsymbol{\beta}^{(j,0)}$ avec des valeurs strictement positives ($\forall j \in \{1,2\}$).
2: **Pour tout** $t = 1, \ldots, T$ **Faire**
3: Mise à jour MU : $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} \odot \mathbf{A}(\mathbf{W}^{(t)})$
4: Mise à jour MU ou Prox : $\forall\, j = 1, 2$, (Prox)

$$\mathbf{H}^{(j,t+1)} \leftarrow \text{prox}_{\tilde{g}_j}\left(\tilde{\mathbf{H}}^{(j)}\right),\ \tilde{\mathbf{H}}^{(j)} = \mathbf{H}^{(j,t)} - \frac{1}{\eta}\nabla f_j(\mathbf{H}^{(j,t)})$$

OU (MU) :

$$\mathbf{H}^{(j,t+1)} \leftarrow \mathbf{H}^{(j,t)} \odot \mathbf{S}(\mathbf{H}^{(j,t)})$$

5: Mise à jour MU : $\forall\, j = 1, 2$, $\boldsymbol{\beta}^{(j,t+1)} \leftarrow \boldsymbol{\beta}^{(j,t)} \odot \mathbf{u}(\boldsymbol{\beta}^{(j,t)})$
6: **Fin pour**
7: **renvoyer** $\mathbf{W} := \mathbf{W}^{(T+1)}$, $\mathbf{H}^{(j)} := \mathbf{H}^{(j,T+1)}$ et $\boldsymbol{\beta}^{(j)} := \boldsymbol{\beta}^{(j,T+1)}$ $(j = 1, 2)$

$\odot$ est l'opérateur de multiplication terme à terme et les valeurs spécifiques de $\mathbf{A}(.)$, $\text{prox}_{\tilde{g}_j}(.)$, $f_j(.)$ $\mathbf{S}(.)$, et $\mathbf{u}(.)$ ont des formes explicites omises ici pour la clarté du propos.

# 3 Applications

## 3.1 Données simulées

L'approche que nous proposons ici a également été évaluée sur des données simulées. Nous avons utilisé le même processus de génération de données que celui décrit dans [15][2] et qui a été utilisé pour évaluer une approche de NMF intégrative non supervisée (iNMF) (voir aussi l'article de comparaison [2] qui utilise ces mêmes données).

En bref, le processus de génération fonctionne en deux temps, le premier correspondant à la génération de données $\mathbf{W}$ et $\mathbf{H}^{(j)}, \forall j \in \{1,2\}$ (ici $n = 50$, $p_1 = 2500$, $p_2 = 400$, $K = 2$) pour lesquelles des signatures typiques des deux groupes sont générées selon une loi $\mathcal{B}eta(2,2) \times 2$ (les autres variables, non informatives, ou les valeurs des variables de signatures pour le groupe complémentaire étant initialisées à la valeur 0). Enfin, dans un deuxième temps, les matrices $\mathbf{X}^{(j)}$ reconstruites à partir de $\mathbf{W}$ et $\mathbf{H}^{(j)}$ sont bruitées.

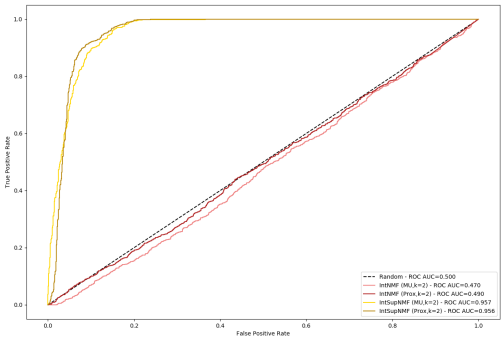
La flexibilité de ce modèle de génération de données nous a permis de tester la NMF intégrative supervisée sur différents aspects, comme la sensibilité au bruit dans les données ou au déséquilibre dans les groupes d'intérêt ou encore dans le nombre de variables caractérisant chacun des groupes. En particulier, nous avons constaté que :

- d'une manière générale, lorsque le niveau de bruit est modéré, la solution fournie par l'approche de résolution proximale extrait des signatures moléculaires directement parcimonieuses, discriminant les individus selon leur groupe d'appartenance. En revanche, lorsque le niveau de bruit dans la génération des données est plus fort, l'approche MU,

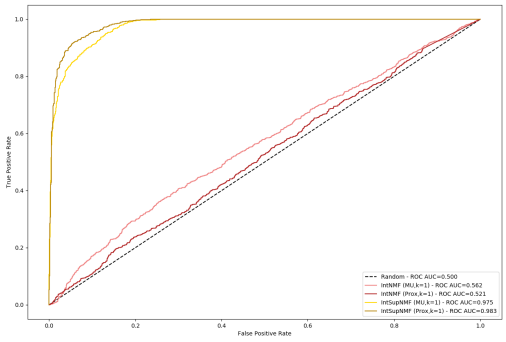
[2]Les scripts associés sont disponibles à l'adresse `https://github.com/yangzi4/iNMF/tree/master`.

qui ne fournit pas de parcimonie exacte dans les signatures extraites, est plus robuste et sélectionne correctement les variables expliquant les groupes ;

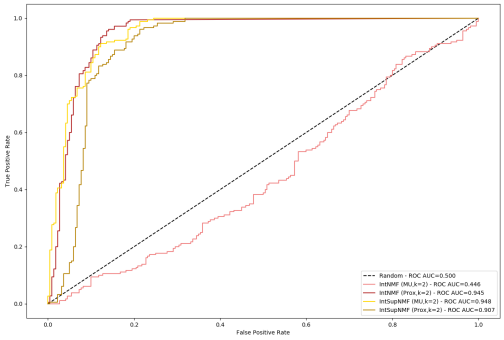
- la partie supervisée dans le terme de reconstruction permet d'améliorer considérablement la qualité des signatures retrouvées, comme illustré sur la Figure 1, et ce particulièrement lorsque le bruit augmente ou que le déséquilibre dans la taille des deux groupes devient important.



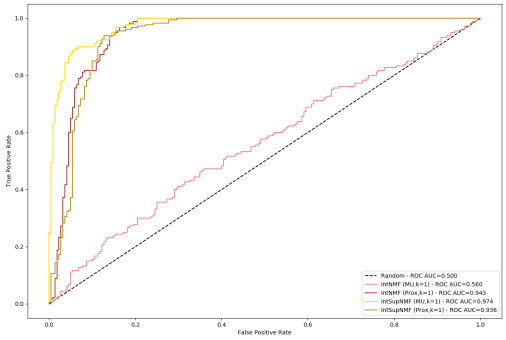

(a) Profile du groupe 1 dans $\mathbf{X}^{(1)}$




(b) Profile du groupe 2 dans $\mathbf{X}^{(1)}$




(c) Profile du groupe 1 dans $\mathbf{X}^{(2)}$




(d) Profile du groupe 2 dans $\mathbf{X}^{(2)}$

FIGURE 1 – Courbes ROC et scores AUC selon la version de la NMF et de l'approche d'optimisation utilisée

Par ailleurs, nous avons comparé notre approche à deux approches d'intégration de données très utilisées, DIABLO [14] (qui est une approche supervisée parcimonieuse basée sur l'analyse canonique des corrélations régularisée) et MOFA [1] qui est une approche non supervisée assez similaire à une Analyse Factorielle Multiple (MFA [5]). Les expériences sont encore en cours mais les premiers résultats semblent montrer que :

- comme la NMF supervisée, DIABLO extrait des signatures moléculaires parcimonieuses et permet bien d'identifier quelques-unes des variables explicatives des groupes dans celles-ci. Toutefois, les signatures obtenues sont généralement trop conservatrices ;

- MOFA n'extrait pas de signatures moléculaires parcimonieuses mais affecte correctement un poids plus important aux variables expliquant les deux groupes, donnant, de ce point de vue, des résultats assez similaires à notre approche.

### 3.2 Etude de la Dermatite Atopique (DA) aux niveaux protéomique et transcriptomique

La NMF supervisée intégrative a également été utilisée pour analyser des données transcriptomiques et protéomiques issues d'une étude sur la Dermatite Atopique (DA) sur peau non lésionnelle. Cette étude a été menée sur $n = 12$ sujets, divisés en deux groupes. Cinq d'entre eux étaient des sujets atteints de DA et les sept autres des sujets sains. Les données finales correspondent ainsi à des données transcriptomiques (issues de la technologie biopuces) contenant l'expression de $p_1 = 22\ 557$ gènes et des données protéomiques contenant la quantification de $p_2 = 281$ protéines.

Sur ces données, la méthode de résolution MU permet de bien discriminer les deux groupes. Les signatures moléculaires sont actuellement à l'étude pour savoir si des éléments connus pour être spécifiques de la DA sont retrouvés.

## 4 Conclusion

Nous avons décrit une approche d'intégration de données adaptée à un problème classique dans les études cliniques dans lesquelles les patients sont souvent classés en groupes d'intérêt biologique. Sur données réelles et simulées, la méthode permet de correctement extraire les signatures biologiques spécifiques des groupes. Les résultats sont en cours d'approfondissement, notamment pour affiner la comparaison avec d'autres approches d'intégration de données mais aussi pour approfondir l'interprétation biologique des signatures extraites pour la DA.

Enfin, notons que la méthode propose un cadre permettant une extension flexible de son utilisation : l'utilisation d'une divergence de Kullback-Leibler à la place de la norme de Fröbenius permettrait de l'adapter à des données fortement non gaussiennes et la modification du terme supervisé permettrait de l'exprimer comme un problème de régression logistique (plutôt que linéaire) ou de régression logistique multiple (pouvant s'adapter à plus de 2 groupes).

## Remerciements

La thèse d'Aurélie Mercadié est financée par les Laboratoires Pierre Fabre et l'ANRT dans le cadre du dispositif CIFRE.

# Bibliographie


[1] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018.

[2] Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier, and Jérémie Becker. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 21(2):541–552, 2020.

[3] Christian Cole, Karin Kroboth, Nicholas J Schurch, Aileen Sandilands, Alexander Sherstnev, Grainne M O'Rega, W H Irwin Watson, Rosemarie M and McLean, Geoffrey J Barton, Alan D Irvine, and Sara J Brown. Filaggrin-stratified transcriptomic analysis of pediatric skin identifies mechanistic pathways in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 134(1):82–91, 2014.

[4] Tara Eicher, Garrett Kinnebrew, Andrew Patt, Kyle Spencer, Kevin Ying, Qin Ma, Raghu Machiraju, and Ewy A. Mathé. Metabolomics and multi-omics integration: a survey of computational methods and resources. *Metabolites*, 10(5):202, 2020.

[5] B. Escofier and J. Pagès. Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 18(1):121–140, 1994.

[6] Pascal Fernsel and Peter Maass. A survey on surrogate approaches to non-negative matrix factorization. *Vietnam Journal of Mathematics*, 46:987–1021, 2018.

[7] Debajyoti Ghosh, Lili Ding, Umasundari Sivaprasad, Esmond Geh, Jocelyn Biagini Myers, Jonathan A Bernstein, Gurjit K Khurana Hershey, and Tesfaye B Mersha. Multiple transcriptome data analysis reveals biologically relevant atopic dermatitis signature genes and pathways. *PLoS One*, 10:e0144316, 2015.

[8] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems (NIPS 2000)*, 13:556–562, 2001.

[9] Hyekyoung Lee and Seungjin Choi. Group nonnegative matrix factorization for EEG classification. In David van Dyk and Max Welling, editors, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 320–327, Clearwater Beach, Florida, USA, 2009. PMLR.

[10] Johannes Leuschner, Maximilian Schmidt, Pascal Fernsel, Delf Lachmund, Tobias Boskamp, and Peter Maass. Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics*, 35:1940–1947, 2019.

[11] Chen Meng, Oana A. Zeleznik, Gerhard G. Thallinger, Bernhard Kuster, Amin M. Gholami, and Aedín C. Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641, 2016.

[12] Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16:85–97, 2015.

[13] Jun-Ichi Sakabe, Koji Kamiya, Hayato Yamaguchi, Shigeki Ikeya, Takahiro Suzuki, Masahiro Aoshima, Kazuki Tatsuno, Toshiharu Fujiyama, Masako Suzuki, Tsuyoshi Yatagai, Taisuke Ito, Toshiyuki Ojima, and Yoshiki Tokura. Proteome analysis of stratum corneum from atopic dermatitis patients by hybrid quadrupole-orbitrap mass spectrometer. *Journal of Allergy and Clinical Immunology*, 134:957–960.e8, 2014.

[14] Amrit Singh, Casey P. Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J. Tebbutt, and Kim-Anh Lê Cao. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics (Oxford, England)*, 35:3055–3062, 2019.

[15] Zi Yang and George Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8, 2016.

# Multi-omic statistical inference of cellular heterogeneity

Hugo Barbot [1] & David Causeur [2] & Yuna Blum [3] & Magali Richard [4]

[1] *IRMAR - UMR CNRS 6625, France, hugo.barbot@institut-agro.fr*
[2] *IRMAR - UMR CNRS 6625, France, david.causeur@institut-agro.fr*
[3] *IGDR - UMR CNRS 6290, France, yuna.blum@univ-rennes.fr*
[4] *TIMC - UMR CNRS 5525, France, magali.richard@univ-grenoble-alpes.fr*

**Résumé.** L'hétérogénéité de la composition en types cellulaires d'échantillons biologiques est un marqueur important de la progression d'une maladie, utile pour son diagnostic. Cette composition cellulaire est cependant difficile à évaluer à partir de profils moléculaires d'un échantillon composite, la contribution de chaque type cellulaire aux signaux observés étant inconnue. La déconvolution cellulaire vise à estimer les proportions des différents types cellulaires à partir de ces profils moléculaires, plusieurs types de données omiques pouvant être utilisés dans cet objectif, tels que l'expression des gènes ou leur taux de méthylation de l'ADN. La déconvolution cellulaire s'appuie sur l'hypothèse que le profil moléculaire de l'échantillon composite peut être approché par une somme pondérée de profils moléculaires spécifiques des mêmes gênes pour chaque type cellulaire considéré, les poids étant les proportions inconnues de ces types cellulaires. La plupart des méthodes statistiques utilisées pour la déconvolution cellulaire sont basées sur des extensions de l'algorithme des moindres carrés ordinaires, sous les contraintes de positivité et de somme à un sur les coefficients du mélange. L'utilisation de cet algorithme suppose implicitement l'indépendance, l'homoscédasticité et la normalité des erreurs résiduelles, conditions sous lesquelles il offre des garanties d'optimalité. Dans le cas présent, chacune de ces trois hypothèses est discutable. D'une part, la nature intrinsèque des données omiques requiert des modèles mieux adaptés à leur sur-dispersion : l'expression des gènes par séquençage de l'ARN est par exemple une donnée de comptage et le taux de méthylation de l'ADN un pourcentage. D'autre part, la structure de dépendance induite par le réseau de régulation des gènes est très forte. Le but de ce travail est de proposer un cadre statistique respectant les caractéristiques inhérentes des données biologiques, et permettant d'intégrer plusieurs types de données omiques.

L'intégration des données multi-omiques pour la déconvolution cellulaire vise à tirer parti de points de vue complémentaires sur l'hétérogénéité cellulaire. Le cadre statistique général que nous proposons est spécialement conçu pour l'intégration de deux types de données omiques fréquemment utilisés pour la déconvolution cellulaire et cités précédemment, l'expression des gènes par séquençage de l'ARN, pour lesquels un modèle contraint de régression binomiale négative est considéré, et le taux de méthylation de l'ADN, utilisant un modèle contraint de régression pour distribution beta. Plusieurs stratégies d'optimisation simultanée sont considérées, basées sur la maximisation sous contrainte de la vraisemblance pondérée, les poids associés aux gènes étant introduits pour renforcer l'influence de certaines combinaisons spécifiques d'expressions et de taux de méthylation de l'ADN, ou sur une sélection de gènes. Une étude comparative de méthodes de déconvolution cellulaire, utilisant conjointement plusieurs types de données omiques ou non, est menée sur des données dites *benchmark*,

utilisant neuf types cellulaires communs dans les études sur le cancer du pancréas (PDAC). Les résultats confirment à la fois le gain de l'utilisation d'une approche multi-omiques et de distributions de probabilités *ad hoc* pour chaque type de données omiques. Finalement, des perspectives d'améliorations fondées sur des modèles de dépendance entre les erreurs d'approximation par les deux types de données omiques pour chaque gène sont présentées.

**Mots-clés.** Déconvolution cellulaire, inférence en grande dimension, Intégration de données multi-omiques, Régression

**Abstract.** Cellular heterogeneity in biological tissues reflects progression of disease state and is therefore useful for improved diagnostic and prognosis. Cellular composition of tissues is however difficult to assess from bulk molecular profiles, with all cells present in the tissue contributing to the recorded signals. Cell deconvolution is a common approach to unravel the heterogeneous molecular profiles observed in bulk tissues, by inferring the underlying relative abundance of individual cell types using one or more omics data, such as RNA-seq gene expressions or DNA methylation rates. So far, cellular deconvolution assumes that bulk omic profiles result from weighted sums of so-called signature cell-specific omic profiles, weights being the unknown proportions of those cell types. Consistently, most statistical methods used for cellular deconvolution are based on extensions of the Ordinary Least Squares (OLS) optimization algorithm, under nonnegativity and sum-to-one constraints on those unknown mixing coefficients. Using OLS implicitly assumes independence, homoscedasticity and normality of the residual errors, conditions under which OLS optimization guarantees optimal estimation. In cellular deconvolution applied to bulk molecular profile, all three assumptions are highly questionable. Indeed, strong violations of those assumptions may be due to the instrinsic nature of omics data, RNA-seq data being overdispersed read counts and DNA methylation rates being percentages for example, or to the dependence structure induced by the gene regulatory network, some key genes being more influent on deconvolution accuracy than others. The goal of this work is to provide a well defined statistical framework that respects the inherent characteristics of biological data for deconvolution, using multi-omic data.

Multi-omic data integration for cellular deconvolution aims at leveraging complementary viewpoints on cellular heterogeneity. The general statistical framework we propose is especially designed for integration of two frequently used omic data types for cell deconvolution mentioned previously, RNA-seq gene expression data, for which a constrained negative binomial regression model is assumed, and DNA methylation rates, using a constrained beta regression model. Many simultaneous optimization strategies are considered, either based on constrained and weighted maximum likelihood, weights being introduced to strengthen the influence of some genes based on their specific combination of signature expressions and DNA-methyation rates, or on gene selection. An extensive comparative study of cell deconvolution performance with leading single or multi-omic methods is conducted on *benchmark* data and using nine cell types commonly found in PDAC (pancreatic cancer). Results confirm both the gain in a multi-omic integration approach and in the use of *ad-hoc* probability distributions for each -omic data type. Additional improvements based on dependence models between approximation errors by the two -omic data types for each gene are finally discussed.

**Keywords.** Cell deconvolution, High-dimensional inference, Multi-omic data integration,

# 1 Introduction to single-omic cell deconvolution

The basic principles of cell deconvolution are introduced hereafter, based on a single genomic profile of gene expressions. For $1 \leq i \leq n$ and $1 \leq j \leq m$, let $y_{ij}$ denote the expression level of gene $j$ for bulk $i$. Let $y_i = (y_{i1}, \ldots, y_{im})'$ denote the complete expression profile for bulk $i$. The signature expressions $x_j = (x_{j1}, \ldots, x_{jK})',\ j = 1, \ldots, m$ of all genes for $K$ cell types of interest is also available. In standard so-called supervised cell deconvolution models, the gene expression profiles are assumed to result from a linear combination of the signature expressions, the mixing coefficients of this linear decomposition being the unknown proportions $\beta_i = (\beta_{1i}, \ldots, \beta_{Ki})'$ of each cell type within each bulk, up to an additive error term. Consistently, most cell deconvolution approaches are variants of the following constrained least-squares minimization issue:

$$(\hat{\beta}_{0i}, \hat{\beta}_i') \quad = \quad \text{argmin}_{(\beta_{0i}, \beta_i')} \sum_{j=1}^{m} (y_{ij} - \beta_{0i} - x_j' \beta_i)^2,$$

where $\beta_{0i}$ is an intercept and the coefficients $\beta_i$ are constrained to lie within the $K-$simplex $\mathcal{S}_K = \left\{ \beta = (\beta_1, \ldots, \beta_K),\ 0 \leq \beta_k \leq 1, \sum_{k=1}^{K} \beta_k = 1 \right\}$.

The nonnegativity and sum-to-one constraints on $\beta$ makes the above constrained optimization issue more challenging than its unconstrained version, which explains the variety of algorithmic solutions available for this task [1]. Moreover, variants of the least-squares objective function have been proposed, aiming for example at more robustness regarding outliers or inspired by popular machine learning methods such as penalized or support vector regression.

Choosing a least-squares type objective function is convenient since well-studied and proven unconstrained minimization algorithms can be used to inspire cell deconvolution methods incorporating nonnegativity and sum-to-one constraints. Moreover, in the maximum-likelihood estimation theory, ordinary least-squares guarantees desirable properties, such as unbiasedness and minimum variance, under the standard assumptions of the linear regression model:

$$y_i \quad = \quad \beta_{0i} \mathbf{1}_m + x \beta_i + \varepsilon_i, \tag{1}$$

where $\mathbf{1}_m$ is the $m-$vector whose entries are all equal to 1, $x$ is the $m \times K$ signature expression matrix whose $j$th row is $x_j$ and $\varepsilon_i$ is an error term assumed to be normally distributed with mean $\mathbf{0}_m$, the $m-$vector whose all entries are zero, and positive variance-covariance matrix $\Sigma = \sigma^2 I_m$, with $\sigma > 0$. In other words, independence, homoscedasticity and normality of the residual error terms are required to guarantee optimality of unconstrained ordinary least squares estimation of the regression coefficients $\beta_i$. In the present situation, all three assumptions are highly questionable. Indeed, whereas in the standard approach of gene

expression data analysis, genes are usually considered as features measured on independent statistical units being different biological samples, in cell deconvolution methods, statistical units are genes and features are bulks. Yet, gene expressions notoriously show different levels of variability and are driven by a gene regulatory network that induces a graph-structured stochastic dependence pattern across genes. Moreover, their distribution is highly skewed, especially when expression data are read counts deduced from RNA-sequencing methods.

For an illustrative purpose, model (1) is fitted to cell deconvolution data in which a profile of $m = 21104$ gene expressions is available on $n = 30$ independent bulks and the signature matrix contains the gene expressions of those $m$ genes in $K = 9$ cell types. The former dataset is generated with the aim of serving as a benchmark reference for comparison of cell deconvolution algorithms. Therefore, it is obtained under a strict control of the true proportions of each of the 9 cell types in the composition of each bulk. Figure 1 displays a heatmap of those true proportions, after a reordering of both cell types and bulks so that similar bulks in terms of cellular composition are grouped in clusters. The plot shows that bulks have different cellular compositions: fibroblasts are obviously dominant in all bulks and the bulks can be divided into two clusters, one with a much larger proportion of classical than basal cancer cells and the other one with more basal than classical cancer cells.

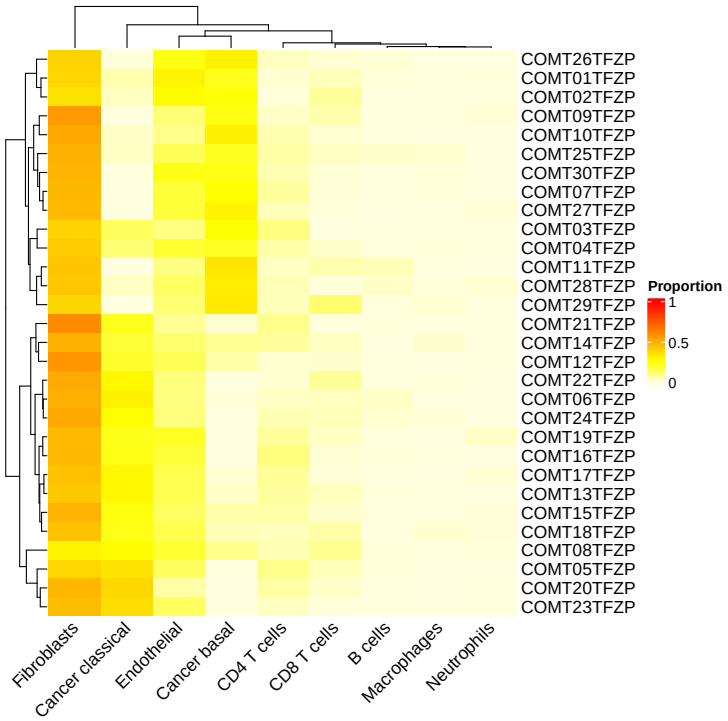



Figure 1: Heatmap of the true proportions of each cell types (columns) within the 30 bulks (rows).

Constrained least-squares approximations of the proportions are now calculated, using

the `R` package `nnls` [2], implementing a nonnegative least-squares estimation algorithm for linear regression models, widely used for cell deconvolution. For each gene, $n = 30$ values of residual errors are calculated as the differences between observed gene expressions in each bulk and linear scores of the signature expressions resulting from the `nnls` algorithm. Figure 2 displays histograms of residual standard deviations (log-transformed for a clearer visualization) and correlations. It shows both a strong heteroscedasticity and dependence across genes, with a strong imbalance between positive and negative correlations and a marked peak of correlations close to 1.

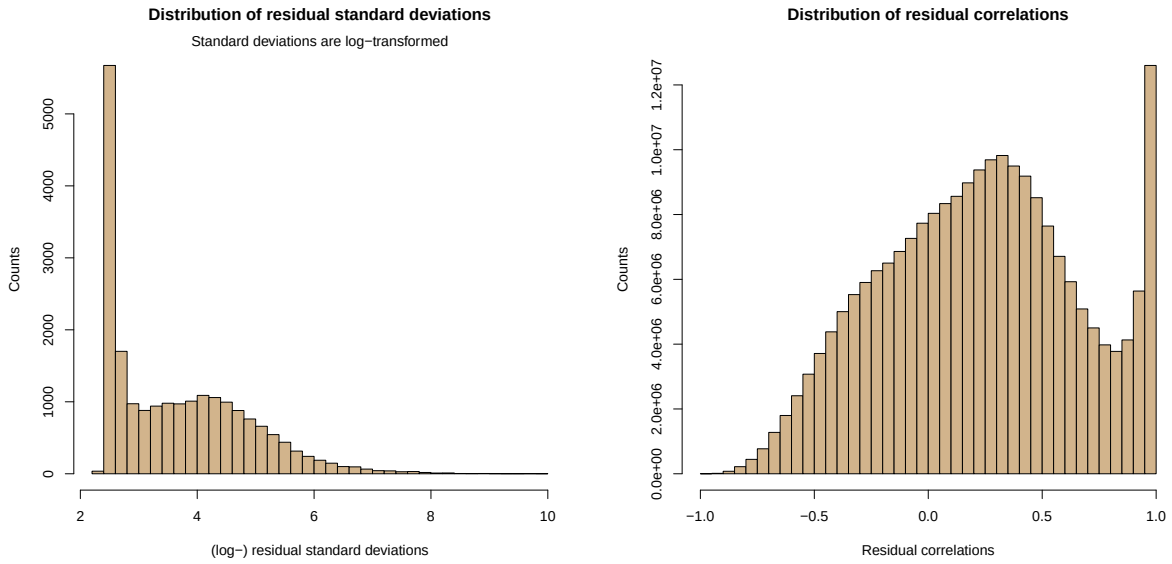



Figure 2: Histogram of log-transformed residual standard deviations (left plot) and residual correlations (right plot).

# 2 A general statistical framework for multi-omic cell deconvolution

Gene expression values obtained using RNA-sequencing technologies are overdispersed read counts between the start and stop codons of each gene. Most of the models used in statistical genomics for analysing such data are based on assumptions of a nonnormal distribution, either Poisson or negative binomial for the most popular.

Similarly, the following constrained negative binomial regression model [3, 4] is now assumed for $Y_{ij}$:

$$\mathbb{P}(Y_{ij} = y_{ij} \mid x_j) \;=\; \frac{\Gamma(y_{ij} + \frac{1}{\alpha_i})}{\Gamma(y_{ij}+1)\Gamma(\frac{1}{\alpha_i})}\Big(\frac{1}{1+\alpha_i\mu_i(x_j)}\Big)^{\frac{1}{\alpha_i}}\Big(\frac{\alpha_i\mu_i(x_j)}{1+\alpha_i\mu_i(x_j)}\Big)^{y_{ij}}, \qquad (2)$$

where $\alpha_i > 0$, $\mu_i(x_j) = \mathbb{E}(Y_{ij} \mid x_j) = \beta_{0i} + \beta_{1i}x_{j1} + \ldots + \beta_{Ki}x_{jK} > 0$, $\beta_{0i} > 0$ and $\beta_i = (\beta_{1i}, \ldots, \beta_{Ki})'$ is the vector of proportions of each cell type, with, for all $k = 1, \ldots, K$, $0 \leq \beta_{ki} \leq 1$ and $\sum_{k=1}^{K} \beta_{ki} = 1$.

Overdispersion with respect to a Poisson regression model that might result from unobserved heterogeneity within the gene expression data is accounted for by parameter $\alpha_i$:

$$\mathrm{Var}(Y_{ij} \mid x_j) \;=\; \mu_i(x_j)(1 + \alpha_i\mu_i(x_j)) > \mu_i(x_j).$$

Under assumption that the gene expression values are independent given the signature expressions, the weighted log-likelihood $\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y)$ of model (2) is given below (index $i$ for the bulk has been omitted):

$$\begin{aligned}\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y) \;=\; & \sum_{j=1}^{m} \omega_{yj} \mathrm{log}\Gamma(y_j + \frac{1}{\alpha}) - \sum_{j=1}^{m} \omega_{yj} \mathrm{log}\Gamma(y_j + 1) - m\mathrm{log}\Gamma(\frac{1}{\alpha}) - \\ & \frac{1}{\alpha} \sum_{j=1}^{m} \omega_{yj} \mathrm{log}\big(1 + \alpha\mu(x_j)\big) + \sum_{j=1}^{m} y_j\omega_{yj} \mathrm{log}\big(\alpha\mu(x_j)\big) - \\ & \sum_{j=1}^{m} y_j \mathrm{log}\big(1 + \alpha\mu(x_j)\big),\end{aligned}$$

where the weights $\omega_y = (\omega_{y1}, \ldots, \omega_{ym})$ are positive, with $\sum_{j=1}^{m} \omega_{yj} = m$. Introducing weights for each gene in the expression of the log-likelihood aims at giving more importance to some influential genes, or even selecting subsets of active genes.

In the PDAC benchmark study introduced above, M-values of methylation levels at more than 800,000 CpG sites over the genome are also available for each of the 30 bulks and correspondingly for the 9 cell types. In order to favor the simultaneous use of DNA methylation and gene expression data in the cell deconvolution task, those methylation rates are aggregated into gene-level measurements by averaging over all values at CpG sites in the promoter region of each gene.

Given the $K-$profile $\tilde{x}_j = (\tilde{x}_{j1}, \ldots, \tilde{x}_{jK})'$ of signature methylation rates for gene $j$ in the cell types of interest, it is now assumed that the M-values $Z_{ij}$ of methylation rates in bulk $i$ are distributed according to a Beta distribution with density [5, 6] :

$$\varphi(z \mid \tilde{x}_j) \;=\; \frac{\Gamma(\phi_i)}{\Gamma(\mu_i(\tilde{x}_j)\phi_i)\Gamma((1 - \mu_i(\tilde{x}_j))\phi_i)} z^{\mu_i(\tilde{x}_j)\phi_i - 1}(1 - z)^{(1-\mu_i(\tilde{x}_j))\phi_i - 1},$$

where $\phi_i > 0$, $\mu_i(\tilde{x}_j) = \mathbb{E}(Z_{ij} \mid \tilde{x}_j) = \tilde{\beta}_{0i} + \beta_{1i}\tilde{x}_{j1} + \ldots + \beta_{pi}\tilde{x}_{jp} > 0$, $\tilde{\beta}_{0i}$ is an intercept parameter and $\beta_i = (\beta_{i1}, \ldots, \beta_{iK})'$ is the vector of proportions of each cell type, with, for all $k = 1, \ldots, K$, $0 \leq \beta_{ik} \leq 1$ and $\sum_{k=1}^{K} \beta_{ik} = 1$.

Under assumption that the M-values are independent given the signature methylation

rates, the weighted log-likelihood $\mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z)$ of the above model is given below:

$$\begin{aligned}\mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z) &= m\log\Gamma(\phi) - \sum_{j=1}^{m} \omega_{zj} \log\Gamma(\mu(\tilde{x}_j)\phi) - \sum_{j=1}^{m} \omega_{zj} \log\Gamma((1-\mu(\tilde{x}_j))\phi) + \\ & \sum_{j=1}^{m} \omega_{zj} \big(\mu(\tilde{x}_j)\phi - 1\big)\log(z_j) + \sum_{j=1}^{m} \omega_{zj} \big((1-\mu(\tilde{x}_j))\phi - 1\big)\big)\log(1-z_j).\end{aligned}$$

As above for the weighted log-likelihood of the gene expression values, gene weights $\omega_z = (\omega_{z1}, \ldots, \omega_{zm})'$ are also introduced here in order to adjust individual gene contributions to the estimation of the cell deconvolution model.

In a multi-omic data integration perspective, we propose a Cyclic Coordinate Descent (CCD) algorithm to optimize $\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y) + \mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z)$ with respect to all parameters, for given weights $\omega_y$ and $\omega_z$. Initial values of the proportion parameters and intercept are obtained by any standard cell deconvolution algorithm, such as `nnls` [2] or `rlm` [7]. Those two algorithms are computationally fast, `rlm` being the most efficient for the time of execution among the three methods handling outlier with bulk data in the benchmark of Avila Cobos *et al* [8]. In order to ensure nonnegativity of estimation, at each update of an estimated proportion parameter, if the marginal maximization of the log-likelihood provides a negative update, then the current value of the proportion parameter is set to zero. Also at each update, the updated vector of proportion parameters is scaled so that it sums to one.

Weights $\omega_y$ and $\omega_z$ can be used to select genes based on their signature profiles of DNA methylation and/or expression values. Indeed, for an illustrative purpose of the former point, a standard hierarchical clustering algorithm applied on the signature profiles of $M-$values provides four clusters showing a gradient of methylation rates for all cell types: in the first cluster, genes have low methylation rates whereas, at the opposite, in the fourth cluster, they have large methylation rates. The weighting strategy considered in the comparative study reported below consists in selecting genes with low methylation rates to fit the cell deconvolution model by setting to zero all weights for genes out of the first cluster, containing 5833 genes.

# 3 A taste of a comparative study

In the present situation where the true proportions $\beta_i$ of each cell type in each bulk are controlled and therefore can be assumed to be known, cell deconvolution methods can be compared using their Mean-Squared-Errors (MSEs) of Estimation, for each bulk over the nine cell types. Part of a large comparative study is reported below, focusing on five cell deconvolution algorithms: first, constrained least-squares approximations of the proportions are calculated for each of the 30 bulks using only the gene expression dataset of the PDAC study, with the `R` package `nnls`, implementing a nonnegative least-squares estimation algorithm for linear regression models, and function `rlm` in the `R` package `MASS`, implementing a robust estimation of a linear regression model using an M-estimator. In the latter case,

negative estimates of the proportions are set to zero and, in both cases, the resulting vector of nonnegative estimated proportions is scaled so that it sums to one. The unweighted log-likelihood of the negative binomial cell deconvolution model (2) is also maximized to provide alternative estimations (`NBR`) of the proportions of cell types using only the gene expression data. M-values in the DNA methylation rates of the PDAC study are introduced in two ways: first by a weighted variant `w-NBR` of the `NBR` algorithm, where, as mentioned previously, the weights are set to zero for genes outside of the first cluster of genes with low methylation rates, and then by combining the former weighting strategy for both omic data types and maximizing the multi-omic log-likelihood (`w-NBR-Beta`).

Figure 3 displays boxplots of an estimation accuracy metric obtained by dividing the MSE for each bulk by the median MSE of the `nnls` method over all bulks. The former relative efficiency measure is introduced in order to figure out the gain with respect to the best OLS-based method in the present study. Additionally, boxplots of the former relative efficiency measures is also provided only for the 14 bulks with more basal than classical cell types. First, it turns out that `nnls` shows better estimation accuracy than `rlm`, and is outperformed by the cell deconvolution approaches we propose. This is especially true when a selection of genes with low methylation rates is introduced in the estimation algorithm and even more when the two -omics data types are simultaneously accounted for in the estimation of the proportions of cell types. The gain in using a multi-omic approach is more obvious when the comparison is restricted to bulks for which the proportion of basal cancer cells exceeds those of basal cancer cells.

# 4 Perspectives

The partial results shown above of a larger comparative study we have conducted based on simulations and on the benchmark datasets of the PDAC study confirms that multi-omic approaches can improve cell deconvolution. The presentation will discuss in which conditions the added value of a multi-omic approach can be expected. Moreover, it will compare a large panel of gene weighting or selection strategies. Finally, the introduction of a dependence model between gene expressions and methylation rates within the statistical framework introduced above will be presented as a possible extension.

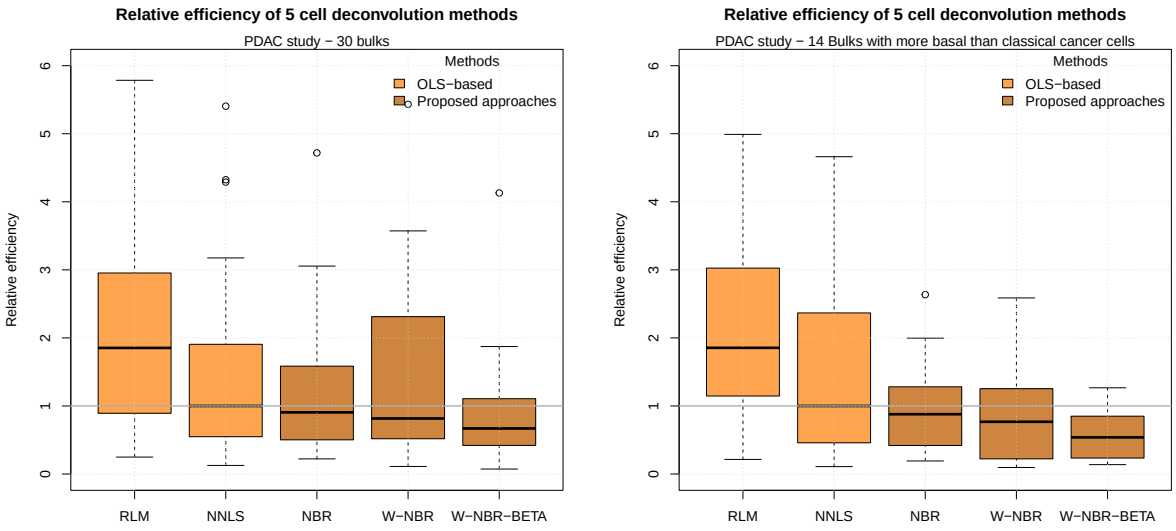



Figure 3: Relative efficiencies of the OLS-based cell deconvolution methods (`rlm` and `nnls`) and three proposed methods: unweighted negative binomial cell deconvolution (`nbr`) based on gene expressions, a weighted negative binomial cell deconvolution (`w-nbr`) also based on gene expressions with weights set to 0 for genes with large methylation rates and a weighted multi-omic (negative binomial + beta regression) cell deconvolution algorithm (`w-nbr-beta`). Left plot: all bulks. Right plot: bulks with more basal than classical cancer cells.

# References


[1] Clémentine Decamps, Alexis Arnaud, Florent Petitprez, et al. DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. *BMC Bioinformatics*, 22(1):473, October 2021.

[2] Katharine M. Mullen and Ivo H. M. van Stokkum. *nnls: The Lawson-Hanson Algorithm for Non-Negative Least Squares (NNLS)*, 2023. R package version 1.5.

[3] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

[4] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[5] Timothy J Triche Jr, Peter W Laird, and Kimberly D Siegmund. Beta regression improves the detection of differential dna methylation for epigenetic epidemiology. *BioRxiv*, page 054643, 2016.

[6] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

[7] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[8] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, et al. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11(1):5650, 2020.

# Analyse différentielle longitudinale des voies métaboliques

Camille Guilmineau [1], Rémi Servien [1], Marie Tremblay-Franco [2,3], Nathalie Vialaneix [4]

[1] *INRAE, Université de Montpellier, LBE, F-11100, Narbonne, France.*
{*camille.guilmineau, remi.servien*}*@inrae.fr*
[2] *INRAE, Université de Toulouse, ENVT, Toxalim, Toulouse F-31027, France.*
[3] *Axiom Platform, MetaToul-MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, Toulouse F-31027, France.*
*marie.tremblay-franco@inrae.fr*
[4] *Université de Toulouse, INRAE, UR MIAT, Castanet-Tolosan F-31326, France.*
*nathalie.vialaneix@inrae.fr*

**Résumé.** La métabolomique permet de décrire le profil métabolique d'un organisme à un instant donné en étudiant les quantités de métabolites, qui sont des molécules de petite taille. Ces métabolites participent au fonctionnement moléculaire des organismes vivants (ou des conglomérats de micro-organismes) au travers de réactions chimiques auxquelles ils participent et les voies métaboliques sont formées par des suites de réactions chimiques impliquant certains métabolites pour une fonction donnée de l'organisme. Ainsi, prendre en compte les voies métaboliques dans les modèles statistiques peut permettre de détecter plus d'effets et de faciliter l'interprétation biologique. Nous nous intéressons ici à l'étude de l'évolution temporelle des métabolites dans un contexte d'analyse différentielle (ou cette évolution est influencée par un facteur d'intérêt) et nous présentons une méthode d'analyse différentielle qui se positionne au niveau de la voie métabolique. Cette méthode comporte deux étapes : la matrice des quantifications des métabolites est d'abord transformée par ACP, puis, un modèle linaire mixte est estimé sur les données transformées. Cette méthode a été appliquée sur des données semi-synthétiques et les résultats ont été comparés à ceux obtenus avec la méthode de référence, l'analyse d'enrichissement. On constate que notre proposition détecte mieux les voies métaboliques différentielles que l'analyse d'enrichissement avec un taux de faux positifs plus faible. La méthode est en cours d'implémentation dans le package R **PHOENICS**.

**Mots-clés.** modèle mixte, données longitudinales, métabolomique, voies métaboliques.

**Abstract.** Metabolomics describes the metabolic profile of an organism at a given time by studying the quantities of metabolites, which are small molecules. These metabolites are involved in the molecular functioning of living organisms (or conglomerates of micro-organisms) through the chemical reactions in which they participate, and metabolic pathways are formed by sequences of chemical reactions involving certain metabolites for a given function in the organism. Thus, taking metabolic pathways into account in statistical models can help detect more effects and facilitate biological interpretation. We are interested here in studying the temporal evolution of metabolites in a differential analysis context (where this evolution is influenced by a factor of interest), and we present a differential analysis method

that is positioned at the pathway level. This methods involves two steps: first, the matrix of metabolite quantifications is transformed by PCA, then a mixed linear model is estimated on the transformed data. This method was applied on semi-synthetic data and the results were compared with those obtained using reference method, namely enrichment analysis. It was found that our proposal detects differential metabolic pathways better than enrichment analysis with a lower false positive rate. This method is currently being implemented in the R package **PHOENICS**.

**Keywords.** mixed model, longitudinal data, metabolomics, metabolic pathways.

# 1 Introduction

La métabolomique consiste à détecter et à quantifier les molécules de petite taille, appelées métabolites, présentes dans des mélanges complexes. Des suites de réactions chimiques se produisent entre les métabolites et sont regroupées dans des voies métaboliques, qui permettent la réalisation d'une fonction utile au système biologique. Des méthodes, présentées dans [Tardivel et al., 2017], permettent de quantifier les métabolites individuellement dans des échantillons biologiques. Ces méthodes haut-débit produisent des données de grande dimension et il est donc nécessaire, pour les analyser, de réduire le nombre de variables largement supérieur au nombre d'individus.

Nous nous intéressons ici au suivi du métabolome, c'est-à-dire à l'évolution de l'ensemble des métabolites d'un système biologique au cours du temps. Dans ce contexte, les données métabolomiques sont acquises à plusieurs dates et sur les mêmes individus, afin d'étudier l'évolution dans le temps du métabolome de ces individus.

Une approche courante pour analyser ce type de données consiste à se baser sur le modèle linéaire mixte, comme proposé par [Martin and Govaerts, 2020]. Ce type de modèle est bien adapté aux données répétées car il ne nécessite pas d'indépendance entre les mesures. Il permet également d'inclure à la fois des effets fixes et aléatoires. Les effets fixes correspondent à des effets contrôlés et d'intérêt, tels que les conditions expérimentales étudiées ou le temps pour les analyse longitudinales. Au contraire, les effets aléatoires représentent des effets non contrôlés, généralement inhérents à la population étudiée, comme la variabilité entre les individus. Cependant, ces approches ne tiennent pas compte des voies métaboliques.

Aussi, en général, les analyses métabolomiques (comme les tests dans les modèles linéaires mixtes décrits ci-dessus), réalisées pour chaque métabolite individuellement, sont post-traitées avec une approche d'analyse d'enrichissement. Elle permet d'étudier si une voie métabolique est enrichie, c'est-à-dire si elle contient significativement plus de métabolites identifiés par l'analyse primaire qu'au hasard.

Nous présentons ici une méthode de modélisation de données métabolomiques longitudinales par voie métabolique. L'analyse par voies métaboliques doit permettre de détecter plus d'effets que l'analyse métabolite par métabolite, car les métabolites d'une voie sont analysés ensemble, permettant de détecter des effets plus faibles qui ne seraient pas identifiés par

l'analyse individuelle des métabolites. Cela doit également faciliter l'interprétation biologique des résultats. L'enjeu est donc d'étendre les approches usuelles afin de construire un modèle mixte basé sur une voie métabolique et non sur un métabolite.

Dans la suite, nous présenterons la méthode proposée dans la section 2. Les données utilisées seront présentées dans la section 3 et la procédure d'évalutation de la méthode sera décrite dans la section 4. Les résultats seront présentés dans la section 5. Enfin, dans la conclusion, nous aborderons des pistes de réflexion et les développements à venir.

# 2 Description de la méthode proposée

La matrice de quantification des métabolites, notée $X$, est de dimension $(n \times T) \times m$, où le nombre total d'observations est égal à $n \times T$, avec $n$ est le nombre d'individus et $T$ le nombre de dates auxquelles ont été mesurées les données, et où $m$ est le nombre de métabolites. On note $p$ le nombre de voies métaboliques contenant ces métabolites. Chaque métabolite appartient à au moins une voie, mais peut aussi être impliqué dans plusieurs voies.

Afin de permettre l'analyse des voies métaboliques, une approche de transformation de la matrice des quantifications des métabolites en une matrice au niveau des voies métaboliques, contenant des scores des voies métaboliques pour chaque individu, a été proposée par [Wieder et al., 2022]. Cependant ce type d'approche ne permet pas la prise en compte des mesures longitudinales.

La méthode proposée est découpée en deux étapes :

1. La première étape consiste à transformer la matrice $X$ des quantifications des métabolites. Pour cela, pour chaque voie métabolique $\mathcal{M}_l$, une ACP est réalisée sur la matrice $Z_l = (X_{ij})_{i=1,\dots,n,\, j\in\mathcal{M}_l}$, la matrice des quantifications des métabolites de la voie métabolique $\mathcal{M}_l$. Les $m_l^*$ premières composantes principales sont sélectionnées par un critère défini préalablement. Nous choisissons ici de sélectionner autant de composantes principales que de facteurs d'intérêts dans le design expérimental. La voie métabolique est alors représentée par les coordonnées des individus sur ces $m_l^*$ composantes principales stockées dans la matrice $A_l$ avec $(n \times T)$ lignes et $m_l^*$ colonnes. Dans la suite, on notera, de manière générique, $a$, une des colonnes d'une des matrices $A_l$, qui correspond donc aux coordonnées sur une des composantes principales.

2. La deuxième étape consiste à estimer un modèle mixte à partir de cette nouvelle matrice, pour décomposer les effets du temps et des conditions expérimentales pour chaque voie métaboliques. Pour cela, on estime le modèle suivant :

$$a = U\beta + D\alpha + \epsilon$$

avec

- $U$ la matrice des $F$ effets fixes, $U = (1|U_1|U_2|...|U_F)$, le temps étant défini comme l'un des effets fixes ;

- $\beta$ le vecteur des paramètres des effets fixes ;
- $D$ la matrice des $R$ effets aléatoires, $D = (D_1|D_2|...|D_R)$. Un effet aléatoire pourra être l'individu sur lequel est réalisée l'observation ;
- $\alpha$ le vecteur des paramètres des effets aléatoires, $\alpha \sim \mathcal{N}(0, \sigma_r^2 \mathbf{I}_{q_r})$ pour un effet aléatoire $r$ ayant $q_r$ niveaux ;
- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_{n \times T})$ les résidus du modèle, supposés i.i.d.

La significativité des effets fixes est testée par ANOVA en comparant le modèle complet à un modèle restreint, de la forme

$$a = U_f \beta_f + D\alpha + \epsilon$$

pour un effet $f$, où $U_f$ est une sous-matrice de la matrice des effets fixes $U$ ne contenant pas l'effet $f$.

Pour une voie métabolique donnée, $\mathcal{M}_l$, cette estimation est répétée pour chacune des colonnes de $A_l$, ce qui conduit à l'obtention de $m_l^*$ $p$-valeurs. La procédure de Simes [Simes, 1986] est utilisée pour agréger, par voie métabolique, ces $m_l^*$ $p$-valeurs : cette procédure contrôle l'erreur de type I de l'hypothèse nulle $\mathrm{H}_0 = \cap_{j=1}^{m_l^*} \mathrm{H}_{0j}$ où $\mathrm{H}_{0j}$ est la nullité de l'effet $f$ pour la $j$-ème colonne de $A_l$. Ainsi, une unique $p$-valeur par voie métabolique est obtenue.

Cette méthode est en cours d'implémentation dans un package R nommé **PHOENICS** que nous présenterons lors des Journées de Statistique.

# 3 Données

Dans le but de tester les capacités de détection de la méthode développée, nous souhaitons utiliser un jeu de données maîtrisé. Pour cela, nous avons créé un jeu de données métabolomiques semi-synthétiques en utilisant la procédure présentée par [Wieder et al., 2022], adaptée aux données longitudinales, et en nous basant sur des données réelles. Ces données proviennent de l'article de [Choo et al., 2017] et sont accessibles dans le dépôt de données métabolomiques MetaboLights (identifiant MTBLS422). Elles contiennent des données de métabolomiques obtenues par résonance magnétique nucléaire (RMN) générées à partir d'échantillons issus d'une étude sur l'effet d'antibiotiques sur des souris. Deux traitements antibiotiques (utilisant de la ciprofloxacin ou du vancomycin-imipenem) ont été comparés à une condition contrôle mais nous nous limiterons à l'utilisation des traitements vancomycin-imipenem et contrôle. Pour chacune des conditions, des mesures ont été réalisées à 3 dates sur 8 souris. Le jeu de données contient ainsi 68 échantillons (4 échantillons sont manquants).

Nous avons traité ces données avec le package R **ASICS** [Lefort et al., 2019] pour obtenir les quantifications des métabolites dans ces échantillons. Les voies métaboliques ont été retrouvées avec le package R **MetaboAnalystR** [Chong and Xia, 2018], qui s'appuie sur la base de données KEGG [Kanehisa and Goto, 2000] spécifique pour l'organisme *Mus musculus* (souris).

Pour créer ces données semi-synthétiques il faut tout d'abord supprimer le signal dans les données réelles. Dans ce but nous avons tout d'abord permuté les échantillons pour les affecter aléatoirement à un groupe (vancomycin-imipenem ou contrôle) et à une date (5.5, 7.5 et 9). Nous avons ensuite choisi $k = 3$ voies métaboliques aléatoirement dans lesquelles nous introduisons une différence entre dates (mais pas entre groupes). Pour cela, les quantifications des métabolites de ces $k$ voies $\{\mathcal{M}_1 \cup \ldots \cup \mathcal{M}_k\}$ sont modifiées en les multipliant par une constante $\alpha_t$, selon la date $t$ :

$$Y_{tij} = X_{tij} \times \alpha_t$$

avec $j \in \{\mathcal{M}_1 \cup \ldots \cup \mathcal{M}_k\}$, $X_{tij}$ la matrice des quantifications et $Y_{tij}$ la matrice des données semi-synthétiques.

Plusieurs scénarios avec différents $\alpha_t$ ont été testés :

$$\text{Scénario 1 :} \quad \begin{cases} \alpha_t = 1 & \text{si } t = 5.5, \\ \alpha_t = 5 & \text{si } t = 7.5, \\ \alpha_t = 10 & \text{si } t = 9, \end{cases}$$

$$\text{Scénario 2 :} \quad \begin{cases} \alpha_t = 1 & \text{si } t = 5.5, \\ \alpha_t = 2 & \text{si } t = 7.5, \\ \alpha_t = 3 & \text{si } t = 9, \end{cases}$$

$$\text{Scénario 3 :} \quad \begin{cases} \alpha_t = 1 & \text{si } t = 5.5, \\ \alpha_t = 1.5 & \text{si } t = 7.5, \\ \alpha_t = 2 & \text{si } t = 9. \end{cases}$$

Afin de permettre l'évaluation de la qualité de la méthode, le tirage aléatoire des 3 voies métaboliques qui sont simulées différentielles a été répété 100 fois.

# 4 Évaluation de la méthode

## 4.1 Comparaison avec les méthodes existantes

Afin d'évaluer notre approche, nous l'avons comparée avec la référence pour l'analyse de voies métaboliques : l'analyse d'enrichissement, qui est basée sur un test exact de Fisher [Wieder et al., 2021]. Une analyse individuelle des métabolites a d'abord été faite en estimant, pour chaque métabolite, un modèle mixte à partir de la matrice des quantifications des métabolites (avec comme effet fixe la date et le traitement et comme effet aléatoire l'individu) puis en testant la significativité des effets fixes. Les métabolites significatifs constituent les métabolites d'intérêt. Pour l'analyse d'enrichissement, un ensemble de métabolites de référence doit également être défini. Il contient généralement l'ensemble des métabolites qui peuvent être détectés dans l'expérimentation. La définition de l'ensemble de référence a un impact important sur les résultats, comme cela a été mis en évidence dans [Wieder et al., 2021], car utiliser un ensemble de référence non spécifique peut mener à un grand nombre de faux positifs. Nous avons testé ici deux ensembles de référence : le premier est constitué de l'ensemble des métabolites de la base de données KEGG, le deuxième est constitué de l'ensemble des métabolites identifiables, ce qui correspond aux métabolites de la base de données du pa-

ckage **ASICS**, utilisé pour identifier les métabolites (soit 180 métabolites). Les deux analyses d'enrichissement ont été réalisées avec le package R **MetaboAnalystR**.

### 4.2 Évaluation de la qualité de la méthode

La méthode a été évaluée à partir des résultats obtenus sur les données semi-synthétiques, avec la méthode que nous proposons et avec les deux tests d'enrichissement. Pour cela, nous avons classé les voies métaboliques en catégories, en fonction de si elles sont significatives et si elles ont été simulées différentielles, comme présenté dans la Table 1. Cependant, les métabolites des voies simulées différentielles peuvent également appartenir à d'autres voies, à cause du chevauchement entre les voies. Les voies chevauchant les voies simulées différentielles ont donc une partie de leurs métabolites qui ont été simulés comme différentiels. Il est donc difficile de conclure pour ces voies car elles ne peuvent pas être considérées complètement comme des voies non différentielles. Elles ont donc été classées dans une catégorie spécifique.

| | Significative | Non significative |
|---|---|---|
| Differentielle | Vrai positifs | Faux négatifs |
| Non differentielle | Faux positifs | Vrai negatifs |
| Non differentielle (chevauchement) | « Faux positifs » (chevauchement) | « Faux négatifs » (chevauchement) |

TABLE 1 : Catégories des voies métaboliques.

Le nombre de voies métaboliques dans chaque catégorie a ensuite été compté sur l'ensemble des répétitions de simulation, pour chacune des méthodes comparées.

## 5 Résultats

Dans le cas des scénarios de simulation 1 et 2, la méthode que nous proposons détecte un plus grand nombre de vrai positifs que les deux tests d'enrichissement. C'est-à-dire que notre méthode retrouve mieux que l'enrichissement les voies métaboliques qui sont effectivement différentielles. Le nombre de faux positifs détectés par notre méthode est également plus faible qu'avec les tests d'enrichissement. Pour les voies qui ont un chevauchement avec les voies différentielles, notre méthode les détecte plus fréquemment significatives que les tests d'enrichissement.

Dans le cas du scénario 3, où les différences sont plus faibles, très peu de voies métaboliques différentielles sont retrouvées par les trois méthodes. Le nombre de faux positifs est également très faible.

# 6 Conclusion

Plusieurs aspects de la méthode nécessiteraient d'être approfondis. Le chevauchement entre les voies métaboliques fait qu'il est difficile de conclure pour ces voies. Il serait intéressant de mieux les étudier et de les caractériser pour comprendre pourquoi certaines voies sont significatives et d'autres non. La calibration du nombre de composantes principales retenues après l'ACP est également un point à approfondir. Ce critère doit permettre de conserver la variabilité dans les données tout en limitant le bruit. Enfin, à plus long terme, nous souhaitons étendre la méthode à l'intégration de données multi-omiques.

D'un point de vue applicatif, la méthode présentée ici sera utilisée pour étudier la formation des photogranules. Les photogranules sont des aggrégats de divers micro-organismes qui présentent des propriétés intéressantes pour le traitement des eaux usées. L'objectif est d'utiliser des données métabolomiques longitudinales afin d'étudier leur développement au cours du temps, ainsi que dans différentes conditions expérimentales. Cela doit permettre d'identifier les voies métaboliques impliquées dans le développement des photogranules et les périodes temporelles importantes.

# 7 Remerciements

Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE45-0036-01.

# Bibliographie

[Chong and Xia, 2018] Chong, J. and Xia, J. (2018). MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics*, 34(24):4313–4314.

[Choo et al., 2017] Choo, J. M., Kanno, T., Zain, N. M. M., Leong, L. E. X., Abell, G. C. J., Keeble, J. E., Bruce, K. D., Mason, A. J., and Rogers, G. B. (2017). Divergent relationships between fecal microbiota and metabolome following distinct antibiotic-induced disruptions. *mSphere*, 2(1):10.1128/msphere.00005–17.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

[Lefort et al., 2019] Lefort, G., Liaubet, L., Canlet, C., Tardivel, P., Père, M.-C., Quesnel, H., Paris, A., Iannuccelli, N., Vialaneix, N., and Servien, R. (2019). ASICS: an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, 35(21):4356–4363.

[Martin and Govaerts, 2020] Martin, M. and Govaerts, B. (2020). LiMM-PCA: combining ASCA+ and linear mixed models to analyse high-dimensional designed data. *Journal of Chemometrics*, 34(6):e3232.

[Simes, 1986] Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[Tardivel et al., 2017] Tardivel, P. J., Canlet, C., Lefort, G., Tremblay-Franco, M., Debrauwer, L., Concordet, D., and Servien, R. (2017). ASICS: an automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. *Metabolomics*, 13(10):109.

[Wieder et al., 2021] Wieder, C., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., Lai, R. P., Bundy, J. G., Jourdan, F., and Ebbels, T. (2021). Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. *PLOS Computational Biology*, 17(9):e1009105.

[Wieder et al., 2022] Wieder, C., Lai, R. P. J., and Ebbels, T. M. D. (2022). Single sample pathway analysis in metabolomics: performance evaluation and application. *BMC Bioinformatics*, 23(1):481.

# Statistique appliquée à l'industrie

# Optimisation Bayésienne en grande dimension: application en physique des réacteurs nucléaires.

Clément Gauchy [1]

[1] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191 Gif-sur-Yvette, France*
*clement.gauchy@cea.fr*

**Résumé.** La quantification des incertitudes (UQ) est une étape cruciale lors de la conception de systèmes industriels complexes, en particulier dans l'industrie nucléaire. Il peut être nécessaire au cours d'une étude d'UQ d'effectuer l'optimisation d'un critère qui est obtenu par une simulation multiphysiques, cette simulation est souvent coûteuse en terme de temps de calcul et de ressources informatiques. Dans cet article, nous étudions un problème jouet consistant à trouver une nappe de puissance optimale d'un réacteur à eau pressurisée (REP) fictif vis à vis d'un critère scalaire, une nappe de puissance étant une distribution spatiale de la puissance à l'intérieur du cœur du réacteur nucléaire. Le critère scalaire pour une nappe de puissance donnée est calculé en utilisant une fonction non linéaire d'une distance entre cette nappe de puissance et une nappe de puissance de référence. Tout d'abord, une réduction de dimension sera appliquée à l'ensemble de données des nappes de puissance. Ensuite, nous comparerons plusieurs algorithmes d'optimisation Bayésienne en termes de performance de regret.

**Mots-clés.** Statistique appliquée, Optimisation Bayésienne, Réduction de dimension.

**Abstract.** Uncertainty Quantification (UQ) is a crucial step when designing complex industrial systems, especially when it comes to nuclear industry. It could be necessary during an UQ study to perform optimization of a criterion which is obtain by a multiphysics simulation, this simulation is often costly in terms of computation time and high-performance computing resources. In this article, we study a toy problem of finding a optimal power map of a fictitious Pressurized Water Reactor (PWR) with respect to a scalar criterion, a power map being the spatial distribution of power inside the reactor core. The optimality score for a given power map is computed using a nonlinear function of a distance measure between the considered power map and a reference power map. First, a dimension reduction will be applied on the power maps dataset. Then, we will benchmark several Bayesian Optimization algorithms in terms of their regret performance.

**Keywords.** Applied statistics, Bayesian Optimization, Dimensionality reduction.

## 1 Introduction

In a Nuclear Power Plant (NPP), power is delivered by the fuel inside the reactor core: the fission reaction inside the fuel realeases heat power that is absorbed by the water from the

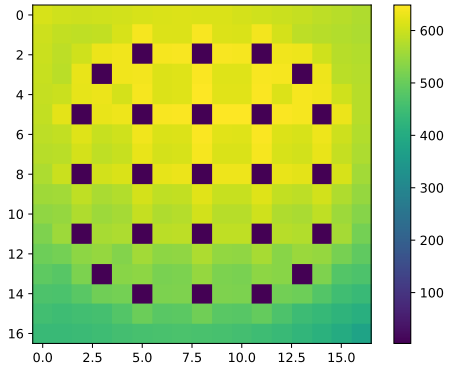
primary circuit. A reactor core in a PWR is composed of fuel assemblies, each assemblies contains 264 fuel rods and 24 tubes that contains the control rods and guide tubes. The fuel assemblies is then arranged in a $17 \times 17$ square lattice. Figure 2.4 of Jacquemain (2015) show a diagram of a fuel assembly. The spatial distribution of heat power inside the reactor core is coined power map and is critical for the exploitation of the reactor and its safety. The increase of computational power since the mid 20th century allows engineers to use numerical simulators to study the fission reaction inside the reactor core and hence computing power maps. During studies for designing the reactor core and assessing the safety of the NPP, several power maps are generated to explore different scenarios during the lifetime of the NPP. For each power map, a multiphysics computation may be done to validate the design and the safety of the NPP. In general, the multiphysics simulation is way costlier in computation time and resources than the simulation of a power map. It is thus preferable to minimize the number of multiphysics simulation to be done. In this paper, we will benchmark several Bayesian optimization (BO) algorithms Wang et al. (2023) with respect to an optimality score, this optimality score will emulate the costly multiphysics computation. The benchmark will be done on a dataset of radial power maps of a mock-up PWR reactor core simulated using the neutronics simulation code APOLLO3 developed in CEA Schneider et al. (2016).

## 2 Description of the data

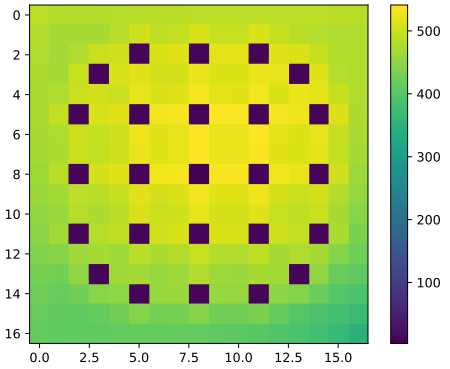
The data used in this paper are $N = 1080$ power maps $(\mathbf{P}_k)_{1\leq k\leq N}$ of a specific fuel assembly of a mock-up PWR reactor core. A power map $\mathbf{P} = (P_{ij})_{1\leq i,j\leq 17}$ is a matrix of dimension 2 composed of $d = 17\times 17 = 289$ values $P_{ij}$ with $P_{ij} \in \mathbb{R}^+$. Hence, we can see that the data are high-dimensional. Figure 1 shows a power map observation as well as descriptive statistics. Note that some locations have always zero values: this corresponds to the positions of the guide tubes, where the neutron absorber rods are introduced to control the nuclear fission reaction. Hence, the number of variables that are informative for our problem for each power map is $17 \times 17 - 25 = 264$. The optimality score for each map is designated by the function $g$ such that for a power map $\mathbf{P}$ we have:

$$g(\mathbf{P}) = \frac{\|\mathbf{P} - \mathbf{P}_{\text{ref}}\|_\infty - m}{\sigma} + \sin\left(\pi \frac{\|\mathbf{P} - \mathbf{P}_{\text{ref}}\|_\infty - m}{\sigma}\right)^2 , \tag{1}$$

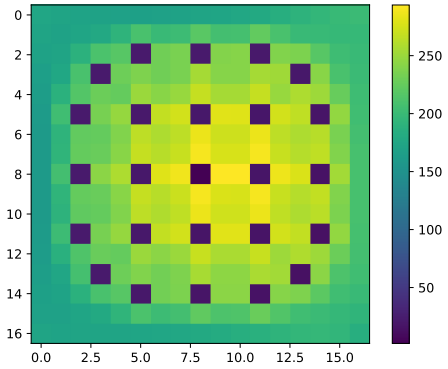
where $m = \frac{1}{N}\sum_{k=1}^{N} \|\mathbf{P}_k - \mathbf{P}_{\text{ref}}\|_\infty$ and $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N} (\|\mathbf{P}_k - \mathbf{P}_{\text{ref}}\|_\infty - m)^2$ and $\mathbf{P}_{\text{ref}}$ is a reference power map consisting of a uniform spatial distribution of the mean total power computed over the dataset, with the total power of a power map defined as the sum of the power in each cell of a power map. The optimality score was designed to emulate a multiphysics simulation of a PWR reactor core with a power map as an input. It is designed to be highly nonlinear in order to reproduce the behavior of chained or coupled numerical simulation between neutronics, thermohydraulics and mechanics.




(a) Example of a power map from the dataset of the mock-up PWR reactor core




(b) Empirical mean power map over the 1080 observations




(c) Empirical standard deviation power map over the 1080 observations

Figure 1: Descriptive statistics of the power maps dataset.

The problem to be solved in this article is to determine $\mathbf{P}_{\max}$ such that:

$$\mathbf{P}_{\max} = \underset{1 \leq k \leq N}{\operatorname{argmax}}\, g(\mathbf{P}_k) \ . \tag{2}$$

The main challenge of (2) is to find $\mathbf{P}_{\max}$ with only the knowledge of $(y_k = g(\mathbf{P}_k))_{k \in \mathcal{I}}$ where $\mathcal{I} \subset \{1, \ldots, N\}$ with $|\mathcal{I}|$ the smaller possible. Indeed, we want to find the solution of our optimization problem with the less number of multiphysics computation possible, here emulated by the optimality criterion $g$. The mathematical tools used to solved this problem will be at first a dimensionality reduction step in order to compress the information given by the 264 power values of each power map and then the use of Bayesian optimization techniques. Bayesian optimization algorithms are designed to optimize a costly objective function with the less number of calls to the function, these methods are adapt to our problem. These two steps are described in the next section.

# 3 Dimensionality reduction

The first step in our study is to reduce the dimensionality of our data, the most popular method for achieving this task is the Principal Component Analysis (PCA) Jackson (1991). Indeed, its popularity is due to its simplicity of computation (it is a convex optimization problem) and its easy intrepretation of the results. Define $\bar{\mathbf{P}} = 1/N \sum_{k=1}^{N} \mathbf{P}_k$ the empirical mean power map. Using linear algebra we can express a power map $\mathbf{P}$ as follow:

$$\mathbf{P} = \bar{\mathbf{P}} + \sum_{k=1}^{N} \alpha^{(k)} \mathbf{P}^{(k)} \ , \tag{3}$$

where $\mathbf{P}^{(k)}$ are the eigenvectors of the covariance matrix of the observed power maps $(\mathbf{P}_k)_{1 \leq k \leq N}$, ranked by decreasing eigenvalues order and

$$\alpha^{(k)} = \langle \mathbf{P}^{(k)}, \mathbf{P} - \bar{\mathbf{P}} \rangle \ ,$$

The main idea of PCA is to truncate the sum in Equation (3) at $M < N$:

$$\mathbf{P} \approx \bar{\mathbf{P}} + \sum_{k=1}^{M} \alpha^{(k)} \mathbf{P}^{(k)} \ , \tag{4}$$

The choice of $M$ is governed by the part of variance explained by the truncated sum:

$$v(M) = 1 - \sum_{k=1}^{M} (\alpha^{(k)})^2 \ , \tag{5}$$

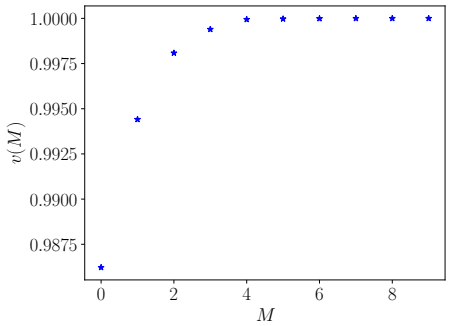



Figure 2: Evolution of the explained variance $v(M)$ with respect to $M$

Figure 2 show the evolution of the explained variance $v(M)$ with respect to $M$. We can observe that the 5 first principal components explain more than 99% of the variance. Hence,

the dimension of data can be drastically reduced from 264 to 5. It is thus possible to express the power maps by a low-dimensional representation thanks to the following application:

$$\varphi : \ \mathbf{P} \mapsto (\alpha^{(i)})_{1 \leq i \leq 5} \tag{6}$$

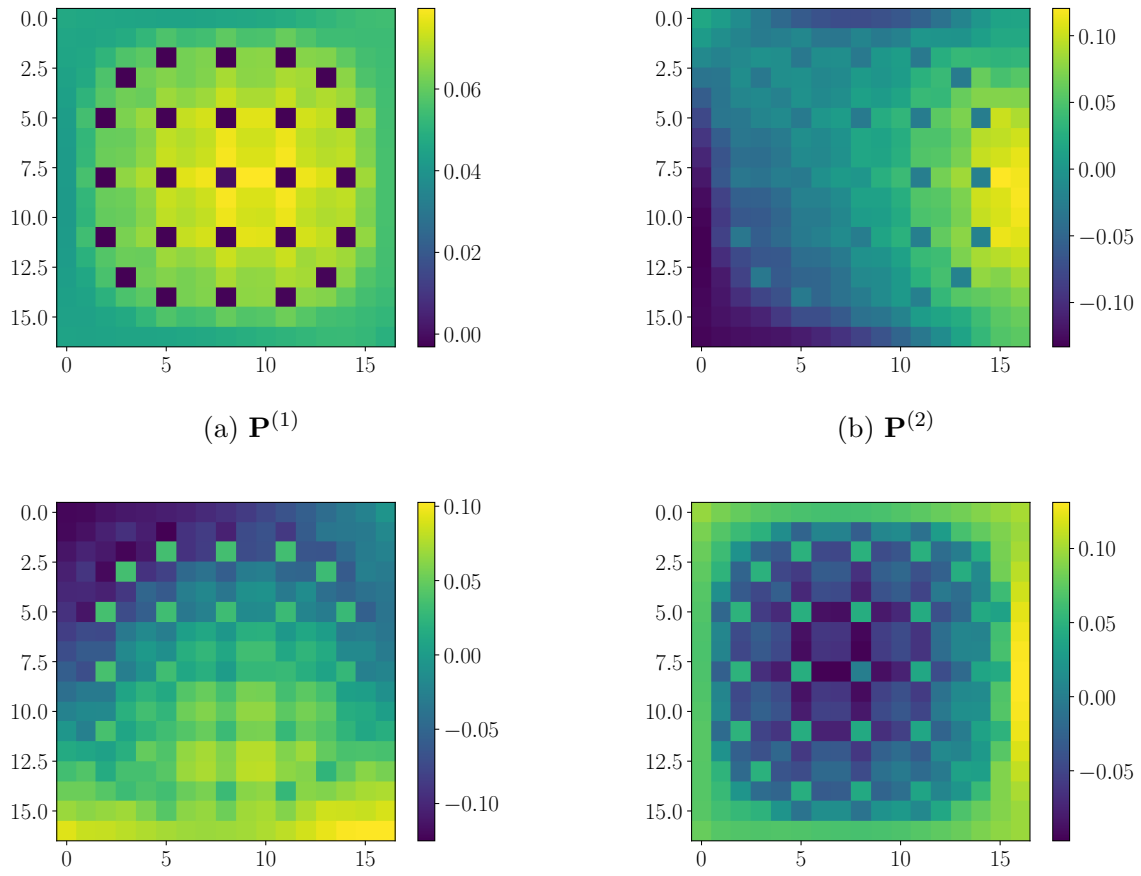
The optimization problem defined in Equation (2) is then reformulate as follow:

$$\mathbf{P}_{\max} = \underset{1 \leq k \leq N}{\operatorname{argmax}} \, \tilde{g}(\boldsymbol{\alpha}_k) \ , \tag{7}$$

where $\boldsymbol{\alpha}_k = \varphi(\mathbf{P}_k) = (\alpha_k^{(1)}, \ldots, \alpha_k^{(M)})$ is the low-dimensional representation of the power map $\mathbf{P}_k$ and $\tilde{g}$ is defined by

$$\tilde{g}(\boldsymbol{\alpha}_k) = g(\bar{\mathbf{P}} + \sum_{k=1}^{M} \alpha_k^{(k)} \mathbf{P}^{(k)}) \ .$$

The main advantage of PCA is its capacity to give visual interpretation of high-dimensional



(a) $\mathbf{P}^{(1)}$ (b) $\mathbf{P}^{(2)}$

(c) $\mathbf{P}^{(3)}$ (d) $\mathbf{P}^{(4)}$

Figure 3: Graphical representation of the four first eigenvectors of the PCA decomposition defined in Equation (4).

data. Figure 3 shows the four first eigenvector of the PCA decomposition. The first eigenvector is associated to high power values in the center of the power map. The second and third eigenvectors are associated to power values on the right and on the bottom of the power map, while the fourth eigenvector concerns the opposite variations of the power between the center and the edges of the power map.

The next section will be dedicated to the benchmark of several Bayesian optimization techniques for solving the optimization problem defined in Equation (7).

# 4 Bayesian optimization

Bayesian optimization aims at optimize objective functions that are costly to evaluate. The main idea is to incorporate prior belief into the objective function in the form of a Gaussian process prior in order to guide sampling of new computations of the costly function, and thus achieving a good compromise between exploration & exploitation of the computational budget. First, we will consider a Gaussian process (GP) $\tilde{G}$ for emulating the optimization criterion $\tilde{g}$. The GP is characterized by a prior mean $\mu(\cdot)$ and a covariance function $c(\cdot,\cdot)$. For $1 \leq n \leq N$, we define $\mathcal{I}_n \subset \{1,\dots,N\}$ a subset of indices of cardinal $n$ and $\mathcal{D}_n = (\boldsymbol{\alpha}_k, \tilde{g}(\boldsymbol{\alpha}_k))_{k\in\mathcal{I}_n}$ the dataset of observations of the optimization criterion. The choice of the next computation is done through the help of the optimization of an acquisition function: these are auxillary functions evaluated using the posterior distribution of the objective function. Several acquistion functions exist in the literature. We will benchmark in this article one of the most popular one, Expected Improvement (EI) Močkus (1975):

$$\begin{aligned} \mathrm{EI}(\boldsymbol{\alpha};\mathcal{D}_n) &= \mathbb{E}_{\tilde{G}}\left[\left(\tilde{G}(\boldsymbol{\alpha}) - \tilde{G}^*\right)_+\right] \\ &= (\mu_n(\boldsymbol{\alpha}) - \tilde{G}^*)\Phi\left(\tfrac{\mu_n(\boldsymbol{\alpha})-\tilde{G}^*}{\sigma_n(\boldsymbol{\alpha})}\right) + \sigma_n(\boldsymbol{\alpha})\Phi\left(\tfrac{\mu_n(\boldsymbol{\alpha})-\tilde{G}^*}{\sigma_n(\boldsymbol{\alpha})}\right) \end{aligned} \tag{8}$$

where $\tilde{G}^* = \mathrm{argmax}_{k\in\mathcal{I}_n}\, \tilde{g}(\boldsymbol{\alpha}_k)$, $\mu_n(\cdot)$ and $\sigma_n(\cdot)$ are respectivly the mean and standard deviation of $\tilde{G}$ conditionally to $\mathcal{D}_n$. Note that the EI acquisition function can be computed analytically without any need of Monte-Carlo sampling. The EI is optimized to determine the next datapoint denoted by $\boldsymbol{\alpha}_{(n+1)}$ such that:

$$\boldsymbol{\alpha}_{(n+1)} = \underset{k\in\{1,\dots,N\}/\mathcal{I}_n}{\mathrm{argmax}}\ \mathrm{EI}(\boldsymbol{\alpha}_k;\mathcal{D}_n)$$

Another classical Bayesian optimization algorithm is denoted by Thompson sampling. The core principles of this method is to sample $\tilde{G}_n \sim (\tilde{G}|\mathcal{D}_n)$ and to solve the optimization problem:

$$\boldsymbol{\alpha}_{(n+1)} = \underset{k\in\{1,\dots,N\}/\mathcal{I}_n}{\mathrm{argmax}}\ \tilde{G}_n(\boldsymbol{\alpha}_k)$$

The next section is dedicated to a benchmark of these two Bayesian optimization method on the power maps dataset.

# 5 Numerical experiments and results

The two Bayesian optimization algorithms presented in Section 4 will be benchmarked on a dataset of 1080 simulated power maps. Also, a random sampling strategy will also be performed in order to compare the Bayesian optimization strategies with respect to a naive strategy of sampling at random the power maps in the dataset. We choose $M = 4$ for the low-dimensional representation of the power maps. Before starting the Bayesian optimization algorithm, 5 power maps are sampled at random in the dataset to train and learn the hyper-paremeters of the Gaussian process, which are then fixed during the rest of the algorithm. 100 replications of each strategies are thus performed to take into account the randomness of the sampling of the 5 first power maps and the one from the sampling of the Gaussian process posterior in the Thompson sampling. The different strategies are quantitatively benchmarked by their regret at step $n$:

$$R_n = g(\mathbf{P}_{\max}) - \max_{k \in \mathcal{I}_n} g(\mathbf{P}_k) \ ,$$

The numerical results of the benchmark is presented in Figure 4. The best strategy given by

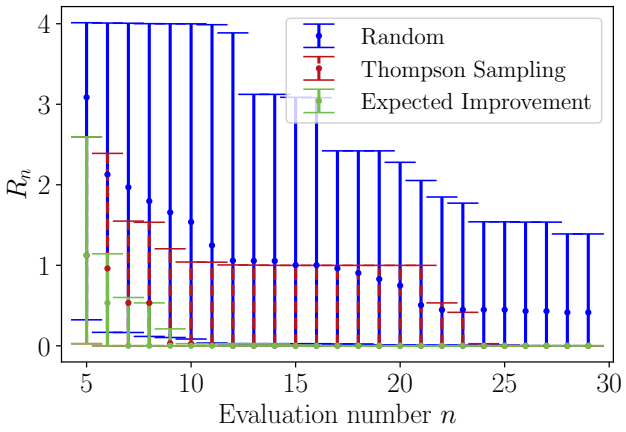



Figure 4: Results from the benchmark of 100 replications of each strategy. The vertical bars represent the interquantiles ranges between the 10% and 90% quantile of the regret $R_n$ at each step $n$.

our results is the EI strategy. Indeed the regret reach the value 0 for all the 100 replications after only 16 computations of the optimization criterion $g$, while Thompson sampling need 25 computations. However, it is important to notice that the two Bayesian optimization algorithms outperformed very well a naive random strategy.

# 6 Conclusion & perspectives

In this article, a toy UQ problem of designing a optimal power map is studied. A optimality criterion is associated to each power map to emulate a costly multiphysics simulation taking a power map as an input. Given the high dimensionality of the power map data, a dimension

reduction procedure by PCA is performed to represent each power map by a low-dimensional vector. After this step, two Bayesian optimization algorithms are benchmarked with respect to their regret values. The Expected Improvement algorithm shows the best performances. A perspective of this work could be to improve the dimensionality reduction by enforcing physics constraints. For instance, the power maps have always positive values, they thus lies in a specific quadrant of $\mathbb{R}^{264}$. More complex physics constraints based on neutronics could be implemented. Another perspective is to implement parallel EI algorithms in the same fashion as in Ginsbourger et al. (2008), Wang et al. (2020). Indeed, parallel computing is now common in industrial practices and thus new Bayesian optimization algorithms needs to be developed for the case when the costly computer simulation is ran in parallel.

# 7 Acknowledgments

The author adress its warmful thanks to Mr. Blaise MATHON for the generation of the power maps dataset in APOLLO3. As well as Dr. Nathan GRENIER, Mr. Thibaut LOPEZ and Mr. Damien RAGUENES for the fruitful discussions, without which this article would not have been possible.

# References

Ginsbourger, D., Le Riche, R. & Carraro, L. (2008), A Multi-points Criterion for Deterministic Parallel Global Optimization based on Gaussian Processes, Technical report.

Jackson, J. (1991), *A user's guide to principal components*, Wiley series in probability and mathematical statistics, Wiley.

Jacquemain, D. (2015), *Nuclear Power Reactor Core Melt Accidents. Current State of Knowledge*, EDP Sciences.

Močkus, J. (1975), On bayesian methods for seeking the extremum, *in* G. I. Marchuk, ed., 'Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 400–404.

Schneider, D., Dolci, F., Gabriel, F., Palau, J.-M., Guillo, M. & Pothet, B. (2016), APOLLO3® CEA/DEN deterministic multi-purpose code for reactor physics analysis, *in* 'PHYSOR 2016 – Unifying Theory and Experiments in the 21st Century', Sun Valley, United States. APOLLO3® is a registered trademark of CEA.

Wang, J., Clark, S. C., Liu, E. & Frazier, P. I. (2020), 'Parallel bayesian global optimization of expensive functions', *Operations Research* **68**(6), 1850–1865.

Wang, X., Jin, Y., Schmitt, S. & Olhofer, M. (2023), 'Recent advances in bayesian optimization', *ACM Comput. Surv.* **55**(13s).

# Noisy radioactivity data analysis using parametric Poisson models

Salima Helali [1], Guillaume Manificat [1], Kévin Galliez [1], Miriam Basso [1] & Maxime Morin [1]

[1] *Institut de Radioprotection et de Sûreté Nucléaire - IRSN, Paris (France), helali.salima@gmail.com*

**Résumé.** En métrologie, l'utilisation des concepts de seuil de décision et de limite de détection pose souvent de nombreux problèmes aux métrologues des laboratoires d'analyse de la radioactivité. Il est habituel de censurer les données lorsqu'il devient difficile de discerner la présence ou l'absence de l'activité, en raison du bruit dans les données de mesure. Cela signifie que si les résultats des mesures sont non significatifs (inférieurs au seuil de décision), l'analyse indique simplement que la valeur réelle (signal) de la radioactivité est inférieure à une certaine limite appelée limite de détection. Ces problèmes sont souvent liés à une mauvaise compréhension des formules de seuil de décision. En outre, la manière de générer un seuil de décision approprié et justifié n'est pas claire dans [ISO 11929(2020)]. Dans le présent document de recherche, nous élaborons une méthode statistique de détermination du seuil de décision le plus puissant. Ensuite, des méthodes d'approches statistiques sont adoptées pour estimer l'espérance et la variance de la radioactivité. L'efficacité et la faisabilité de ces approches sont corroborées par des applications sur des ensembles de données réelles de l'IRSN.

**Mots-clés.** Seuil de décision, Test du rapport de vraisemblance, Radioactivité, Données de bruit, Paramètres de nuisance.

**Abstract.** In metrology, the use of the concepts of decision limit and detection limit often poses many problems for metrologists in radioactivity analysis laboratories. It is usual to censor data when it becomes difficult to discern the presence or absence of activity, due to noise in the measurement data. This means that if the measurement results are insignificant (below the decision threshold), the analysis simply indicates that the actual value (signal) of radioactivity is below a certain limit called the detection limit. These problems are often due to a misunderstanding of the decision threshold formulas. In addition, it is not clear how to generate an appropriate and justified decision threshold in [ISO 11929(2020)]. In this research paper, we develop a statistical method for determining the most powerful decision threshold. Next, methods of statistical approaches are adopted to estimate the expectation and the variance of radioactivity. The effectiveness and feasibility of these approaches are corroborated by applications on IRSN data sets.

**Keywords.** Decision threshold, Likelihood ratio test, Predictive density, Radioactivity, Noise data, Nuisance parameters.

# 1 Introduction

Radioactivity measurement relies on characteristic limits (decision thresholds and detection limits). This commonly indicates that measurement results below the decision threshold (DT) are left censored and are considered as useless. Due to ever-lower levels of environmental activity, the number of radiological analyses for which metrologists are unable to provide significant results is increasing. Current standards steadily propose uniquely DT formulas without much justification. For instance, the DT, as defined by the standard [ISO 11929(2020)], is calculated as

$$y^\star = k_{1-\alpha} w u(n_n = 0),$$

where $k_{1-\alpha}$ denotes the quantile of the probability density of the measurement results $y$ with a null parameter that exceeds the DT with the probability $\alpha$, $w$ refers to the value of a conversion factor and $u(n_n = 0)$ stands for the null measurement uncertainty of the net indication $n_n$. It is unclear in the standard [ISO 11929(2020)] how the DT is calculated and under which statistical test it is obtained. Currently, there is no clear method for the determination of DT of the radioactivity in literature. There is equally an imperious need in metrology for a clear method to determine an optimal DT due to ever-lower levels of environmental activity. The Institute for Radiological Protection and Nuclear Safety (IRSN) is thus highly interested in improving DT for radioactivity analyses. The aim of this short communication is to present an efficient statistical method capable of providing an optimal DT and study the statistical properties of the radioactivity.

# 2 Model and notations

In the field of metrology, in order to quantify the activity contained in various samples, and before the actual measurement of a specific sample is undertaken a calibration procedure is done. In measurement technology and metrology, calibration is the comparison of measurement values delivered by a device under test with those of a calibration standard of known accuracy. The formal definition of calibration by the International Bureau of Weights and Measures (BIPM) is the following: "Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication."

For the measurements considered here, the calibration process leads us to the following counting procedure: for each sample $A$, two measurements of equal duration are taken: namely a blank measurement $C_{blank}$ and a gross measurement $C_{gross}$, where $C_{blank}$ is as similar as possible to $C_{gross}$ but in condition that ensures the absence of the signal $S$. According to [ISO 11929(2020)], the Noisy radioactivity data are defined as follows

$$A = Y \left( C_{gross} - C_{blank} \right),$$

where

- $C_{blank} := C_{bl}$: is a random variable that models the noise of the counter: $C_{bl} \sim \mathcal{P}(\lambda)$.
- $C_{gross} := C_{gr} = S + B'$: is a random variable that is equal to the sum of two independent random variables, measured by counter, with $S$ being the random variable of interest that models the signal and $B'$ being a random variable that models the noise of the measurement of $C_{gr}$. $S \sim \mathcal{P}(\theta)$, $B' \sim \mathcal{P}(\lambda')$, then $C_{gr} \sim \mathcal{P}(\mu := \theta + \lambda')$.
- $Y$: is a Gaussian random variable that models the conventional calibration factor used in metrology, which maps the linear relation between the measured activity from a reference measurement system and the the measured activity coming from the counter and depends on the time duration and the sample volume.

In section 4, we consider the measurements of radionuclides in water samples as an example of this model of data. We will use the following notations:

- $X = C_{gr} - C_{bl}$: is the net count.
- $DT$: is the decision threshold: critical level of the hypothesis statistical test of $\theta$: $H_0 : \theta = 0$ (no signal) / $H_1 : \theta > 0$ (signal).
- $DL$: is the detection limit: the largest true value $\theta$ that would have a non-negligible probability of being considered insignificant.
- $R := X\mathbb{1}_{\{X \geq DT\}} + DL\mathbb{1}_{\{X < DT\}}$.
- $D := YR$: models the radioactivity used by IRSN which has been so far based on [ISO 11929(2020)]. There is a loss of information because non significant results are left censored.
- $n$: is the size of data.
- $\phi$: is the Gaussian distribution of $\mathcal{N}(0, 1)$.

Let $(A_1, \ldots, A_n)$ and $(D_1, \ldots, D_n)$ be two samples of independent identically distributed (i.i.d.) of random variables defined as follows:

$$\left( A_i = Y_i X_i = \frac{f_i}{\epsilon_i t_i V_i} (C_{gr,i} - C_{bl,i}) \right)_{1 \leq i \leq n},$$

$$vs$$

$$\left( D_i = Y_i R_i = \frac{f_i}{\epsilon_i t_i} \left( X_i \mathbb{1}_{\{X_i > DT_i\}} + DL_i \mathbb{1}_{\{X_i \leq DT_i\}} \right) \right)_{1 \leq i \leq n},$$

where $X_i = \dfrac{1}{N_i} \displaystyle\sum_{k=1}^{N_i} X_i^k$, $DT_i$, $DL_i$ are constructed by $X_i^1, \ldots X_i^{N_i}$ and $N_i$ is the number of repetition of the measurements of the observed point data $X_i$. Noting that $Y_i$ and $C_{bl,i}$

are constants for some $i$, all the parameters $\theta$, $\lambda$, $\lambda'$, $\mu$ are unknown and $\theta$ is the interest parameter. Throughout this research work, the following assumption will be considered:

$$\mathcal{A} : \lambda = \lambda'.$$

This assumption is usually justified by the fact that metrologists put a great emphasis on ensuring that the measurements are undertaken in very similar physical conditions (temperature, background radioactivity, apparatus, etc).

# 3 Main results

Our first result is the following proposition which gives the optimal $DT_i$ for $i = 1 \dots n$. A $DT_i$ is constructed based on the repeated measurements $\left(C^j_{bl,i}, C^j_{gr,i}\right)_{1 \leq j \leq N_i}$ and the following pair of hypotheses test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta > 0,$$

with a fixed risk $\alpha$.

**Proposition 1 (DT of $\theta$)** *Under the hypothesis test $H_0 - H_1$ and the assumption $\mathcal{A}$, an optimal DT of the parameter $\theta$ for $i = 1 \dots n$, denoted by $DT_i$, is generated by the resolution of the following equation:*

$$\alpha \quad = \quad I_{1/2}(DT_i + 1, \overline{C_{bl,i}} + 1),$$

*where $I$ is the regularized beta incomplete function. An approximation of $DT_i$ is determined by*

$$DT_i = \frac{k^2_{1-\alpha} + \sqrt{k^4_{1-\alpha} + 8k^2_{1-\alpha}\overline{C_{bl,i}}}}{2},$$

*where $k_{1-\alpha} = \phi^{-1}(1-\alpha)$.*

The proof of Proposition 1 relies upon the predictive likelihood ratio test devoted to detect the presence of the signal. Noting that this test is uniformly the most powerful based on Neyman-Pearson Lemma. The following proposition specifies a decision threshold $DT_i$ for $i = 1 \dots n$ in the case where $C_{bl}$ is large. We consider that $C_{bl}$ is large when $C_{bl} > 10$.

**Proposition 2 (DT when $C_{bl}$ is large)** *Consider that $C_{bl}$ is large. Under the hypothesis test $H_0 - H_1$ and the assumption $\mathcal{A}$ , an optimal DT of the parameter $\theta$ for $i = 1 \dots n$, is obtained by the resolution of the following equation:*

$$\alpha \quad = \quad 1 - \frac{\Gamma(\lfloor DT_i + 1 \rfloor, 2\overline{C_{bl,i}})}{\lfloor DT_i \rfloor},$$

*where $\Gamma(.,.)$ is the upper incomplete gamma function and $\lfloor . \rfloor$ is the floor function.*

We equally find $DT_i$ based on numerical Newton Raphson method in order to solve the previous equation using package *rootSolve* with the $R$ software or by an approximation of the incomplete beta function. A confidence interval of the signal $S$ is of the form: $\left[x_i - DT_i\frac{\sqrt{X_i}}{N_i}, x_i + DT_i\frac{\sqrt{X_i}}{N_i}\right]$. Finally, the following proposition provides global estimators of the expectation and the variance of the radioactivity.

**Proposition 3 (Estimation of the radioactivity)** *If $X$ and $Y$ are independent, then the considered estimators of the expectation and the variance are indicated by $\mu_{A,n}$ and $\sigma^2_{A,n}$ as follows*

$$\mu_{A,n} = \overline{X} \times \overline{Y},$$

$$\text{and } \sigma^2_{A,n} = \overline{Y}^2\sigma^2_{X,n} + \overline{X}^2\sigma^2_{Y,n},$$

*where* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\sigma^2_{X,n} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *and* $\sigma^2_{Y,n} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

The proof of Proposition 3 is based on the bi-variate first order Taylor expansion of the function $g(X, Y) := XY$ at the point $a := (\mu_X, \mu_Y)$, where $\mu_X$ and $\mu_Y$ are the expectation of $x$ and $Y$ respectively.

# 4 Numerical studies

In our research work, in order to illustrate the proposed theoretical results, we used a data set of measurements of radionuclides released into water. In particular, we considered the data set consisting of $n = 360$ observations within the period $2019 - 2022$ of the following variables

- $S$: the measurement of tritium in environment (mainly rivers in France) by liquid scintillation counting by Bq/L.
- $C_{bl}$: pure water (deep water).
- $C_{gr} = S + B'$, where $B' \sim C_{bl}$.
- $Y$: calibration factor.

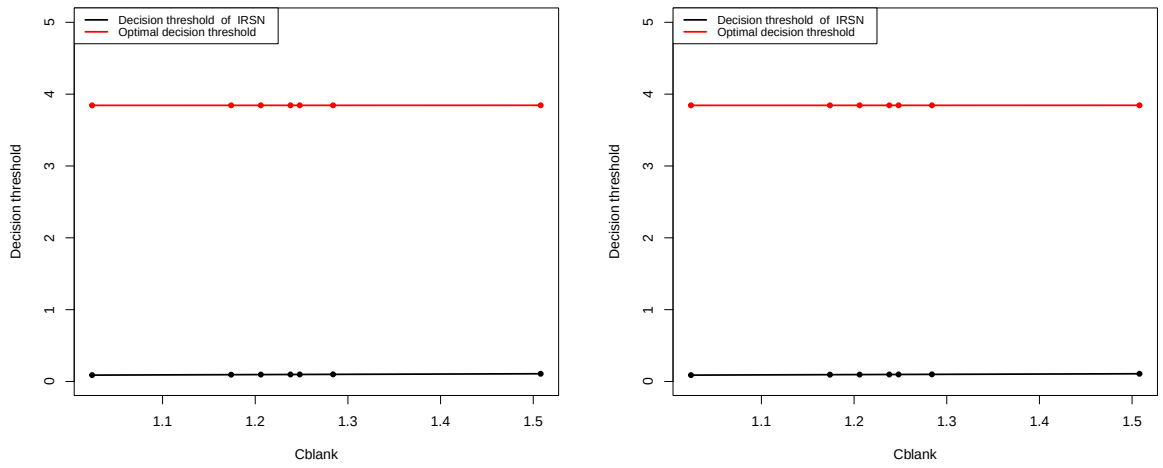



Figure 1: Results for DT of the count part for the radioactivity data in 2019 (left) and 2020 (right)

| | Result of IRSN | Proposed method |
|---|---|---|
| Activity | 6.08 | 6.08 |
| DT | 0.735 | 3.84 |
| Confidence interval | $[6.08 \pm 0.94] = [5.13, 7.02]$ | $[6.08 \pm 1.89] = [4.18, 7.97]$ |
| Conclusion | $A > DT$ | $A > DT$ |

Table 1: Example 1

| | Result of IRSN | Proposed method |
|---|---|---|
| Activity | 1.08 | 1.08 |
| DT | 0.737 | 3.32 |
| Confidence interval | $[1.08 \pm 1.38] = [-0.30, 2.38]$ | $[1.08 \pm 1.22] = [-0.14, 2.30]$ |
| Conclusion | $A > DT$ | $A < DT$ |

Table 2: Example 2

Figures 1 and 2 demonstrates that the proposed DT is higher than the DT defined in [ISO 11929(2020)], which implies that certain radioactivity values are declared significant with the standard when in fact they are not (see for instance Tables 1 and 2). With reference to Tables 1 and 2, we infer that the confidence interval of [ISO 11929(2020)] converge to the proposed confidence interval which is based on the proposed DT. Figure 3 shows that ratio of theoretical false positives and the observed false positives $\alpha_{thoeretical}/\alpha_{observed}$ is closer to 1 especially when $C_{blank}$ is large, which implies the proposed decision threshold is optimal, as it is theoretically expected.

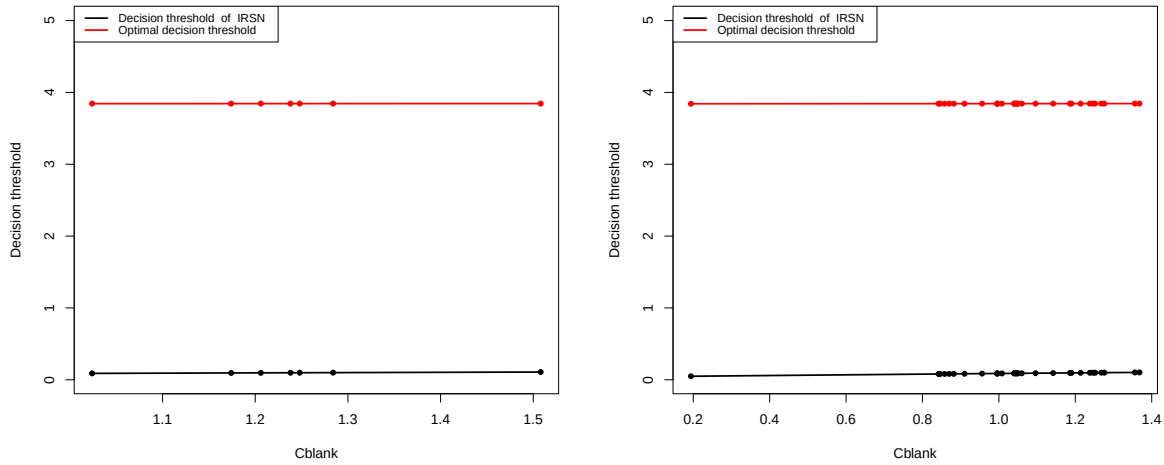



Figure 2: Results for DT of the count part for the radioactivity data in 2021 (left) and 2022 (right)

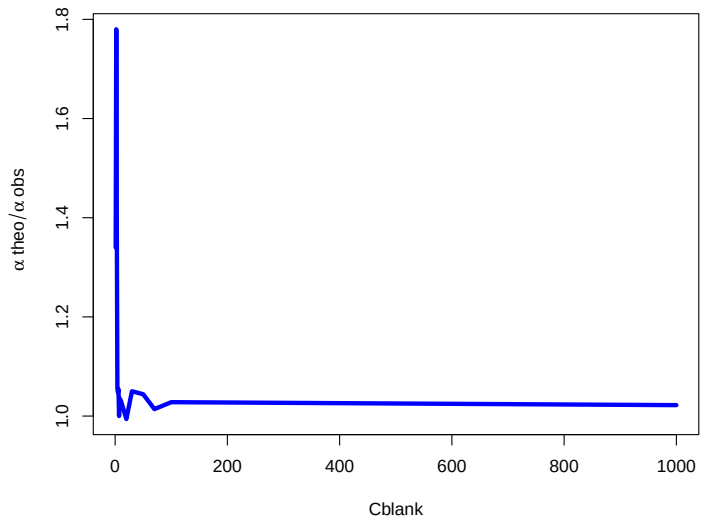



Figure 3: The ratio of theoretical false positives and the observed false positives $\frac{\alpha_{theo}}{\alpha_{obs}}$.

# 5 Conclusion

In this research paper, we proposed an optimal decision threshold based on Neyman-Pearson Lemma in order to detect the signal of the radioactivity. Theoretical results related to the model of the radioactivity; in particular to the expectation, the variance and the confidence interval of the activity signal, are presented. The application of our method was illustrated with the study on real datasets of tritium in water. In this respect, we would assert that this synthesis can be regarded as a preliminary study for further investigations on the decision threshold of the activity signal. Indeed, our work provides a theoretical foundation of the optimal decision theshold in the case of the alpha/beta/gamma spectrometry.

# References

[Barbour et al(1992)] Barbour, A. D., Holst, L., and Janson, S., (1992). Poisson approximation. *Oxford University Press.*,

[Lehmann et al(2005)] Lehmann, E. L., Romano, J. P., and Casella, G. , (2005). Testing statistical hypotheses. *New York: springer.*, **3**.

[ISO 11929(2020)] ISO 11929, (2020). Determination of the characteristic limits (decision threshold, detection limit and limit of the confidence interval) for measurements of ionizing radiations - Fundamentals and Application. *International Organization for Standardization, Geneva.*

[Rosenblatt(1956)] Rosenblatt, M.,, (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The annals of mathematical statistics*, **27** 832—83.

# Forecasting Net Load in France The EDF Data Challenge

Eloi Campagne[1] Yvenn Amara-Ouali[2] Yannig Goude[3] Argyris Kalogeratos[4]

[1,4] *Centre Borelli, Université Paris-Saclay, ENS Paris-Saclay*
[1,2,3] *OSIRIS, EDF R&D*
{eloi.campagne, argyris.kalogeratos}@ens-paris-saclay.fr
{yvenn.amara-ouali, yannig.goude}@edf.fr

**Résumé.** Cet article présente un Data Challenge axé sur l'amélioration de la prévision de la demande nette d'électricité à court terme dans le contexte d'un réseau électrique décentralisé. L'intégration croissante des sources d'énergie renouvelables et les incertitudes liées à leurs fluctuations soulignent la nécessité d'une prévision précise de la demande. Les participants du challenge ont pour objectif de prévoir la consommation nette, c'est à dire la consommation totale moins la production renouvelable – solaire et éolienne –, en France, en utilisant des données disponibles à différentes résolutions géographiques (régionales et nationales), ainsi que des données de prix de l'électricité. Ce papier présente d'abord le contexte de la prévision de demande nette en électricité, puis présente les solutions apportées en amont du challenge.

**Mots-clés.** Data Challenge, Apprentissage automatique, Modélisation statistique, Production d'énergie renouvelable.

**Abstract.** This article presents a Data Challenge focused on improving short-term electricity net demand forecasting in the context of a decentralized power grid. The complexities arising from the increasing integration of renewable energy sources, and the uncertainties associated with their fluctuations, underline the need for accurate demand forecasting. Challenge participants aim to forecast net consumption, i.e. total consumption minus renewable production – solar and wind –, in France, using data available at different geographical resolutions (regional and national), as well as electricity price data. This paper first presents the context of the problem of net electricity demand forecasting, and then presents some of the interesting solutions produced for the challenge.

**Keywords.** Data Challenge, Machine Learning, Statistical Modelling, Renewable Energy Production.

## 1 Context and motivations

The effective operation of the electrical system relies on maintaining a balance between electricity supply and demand. Since electricity cannot be stored, its production needs to be constantly adjusted to match consumption. Providing accurate forecasts for short-term electricity demand is therefore crucial for all participants in the energy market. The shift towards

a decentralized electricity network introduces new uncertainties, which pose additional challenges for demand forecasting. The increasing contribution of renewable energy sources like solar and wind power, brings fluctuations and intermittency to the electricity market. These fluctuations and intermittency occur at various spatial scales due to the presence of wind farms and photovoltaic power plants. The Covid crisis, along with the current economic downturn, add further complexity to forecasting due to the non-stationarity in consumption patterns (Alasali et al., 2021). The availability of new geolocalized data and individual electricity consumption data can be exploited by models that are able to take advantage of additional information and help in minimizing forecast uncertainty (Obst et al., 2021, de Vilmarest and Goude, 2021). Furthermore, recent advancements in adaptive forecasting algorithms have demonstrated improvements in forecasting quality, particularly for aggregate load forecasting (Brégère and Huard, 2022, Antoniadis et al., 2022). However, most existing research overlooks the valuable information available at different spatial or relational scales. The aim of the EDF's FNL Challenge is therefore to focus on methods that can take into account geolocalized data to forecast net demand over France: in particular, the aim is to model renewable electricity production as accurately as possible.

With the announcement of the challenge and the call for participation, the organisers provided a starter kit containing pre-filled code notebooks designed to help participants familiarise themselves with the context of the problem. The starter kit includes a document that motivates the challenge, provides background information and describes the technical aspects of the forecasting problem. It also includes the datasets of interest, code for running a typical pipeline to handle the task (i.e. data loading and visualisation), and a number of basic models that provide a performance benchmark for the problem, with ready-to-use optimisation modules. Section 2 describes all of these elements, followed by Section 3 which presents the solutions produced in advance of the challenge.

# 2 Net Load Forecasting in France

## 2.1 Datasets

Two consumption datasets are available for the challenge: a regional dataset $\mathcal{D}_r$ and a national dataset $\mathcal{D}_n$. Each of them includes meteorological information (temperature, wind, cloud cover, etc.) and calendar information (time of day, type of day, holiday, etc.), enabling net demand to be forecast. A price dataset $\mathcal{D}_p$ is also available, with information on spot prices. For the challenge, the initial training period $\mathcal{T}_{\text{init}}^{\mathbf{tr}}$ runs from 2016-06-01 00:00 to 2021-06-01 23:30. The Covid containment periods (from 2020-03 to 2020-05-10, then from 2020-10-30 to 2020-12-15, and finally from 2021-04-03 to 2021-05-03) have been invalidated. The initial test period $\mathcal{T}_{\text{init}}^{\mathbf{te}}$ runs from 2021-06-02 00:00 to 2022-06-01 00:00:00. As the test period is hectic, we have retained only the last week of each month and the entire month of May 2022 in the test dataset. The remaining weeks have been added to the training dataset, we denote those weeks by $\Delta\mathcal{T}^{\mathbf{tr}}$. So, the final training period $\mathcal{T}_{\text{fin}}^{\mathbf{tr}}$ and test period $\mathcal{T}_{\text{fin}}^{\mathbf{te}}$ can be expressed as $\mathcal{T}_{\text{fin}}^{\mathbf{tr}} = \mathcal{T}_{\text{init}}^{\mathbf{tr}} \cup \Delta\mathcal{T}^{\mathbf{tr}}$ and $\mathcal{T}_{\text{fin}}^{\mathbf{te}} = \mathcal{T}_{\text{init}}^{\mathbf{te}} \setminus \Delta\mathcal{T}^{\mathbf{tr}}$ (Figure 1).

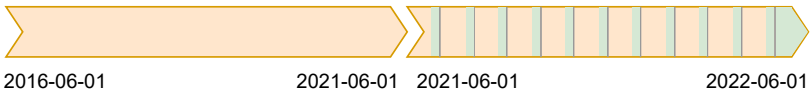



Figure 1: Training and test periods. Orange corresponds to $\mathcal{T}_{\text{fin}}^{\mathbf{tr}}$ and green to $\mathcal{T}_{\text{fin}}^{\mathbf{te}}$.

**Regional dataset.** In this dataset, 12 administrative regions of France are considered: Hauts de France, Normandie, Ile de France, Grand Est, Bretagne, Pays de la Loire, Centre Val de Loire, Bourgogne, Franche Comte, Nouvelle Aquitaine, Auvergne Rhone Alpes, Occitanie, and Provence Alpes Cote d'Azur. Note that Corse is not part of the study. In each region, there is a number of weather stations, between 1 and 5, and the meteorological variables are aggregated with a weighted average over these stations. Finally, a linear interpolation is used to obtain half-hourly data.

**National dataset.** The data in $\mathcal{D}_n$ are of the same type as those in $\mathcal{D}_r$, and differ only in their geographical resolution.


Figure 2: Map of the French mainland with its 12 administrative regions, and 32 weather stations appearing as black dots.

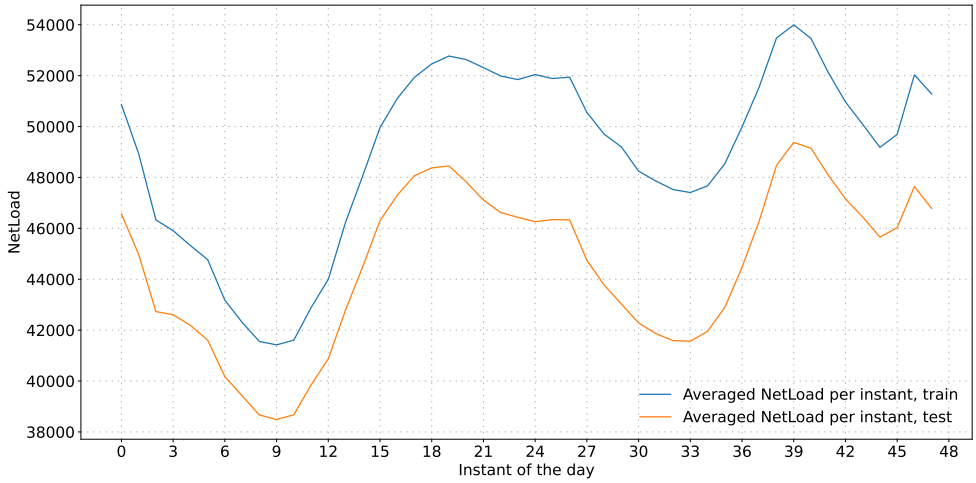



Figure 3: Averaged Net Load over time, for $\mathcal{T}_{\text{fin}}^{\mathbf{tr}}$ and $\mathcal{T}_{\text{fin}}^{\mathbf{te}}$.

**Price dataset.** The price data in $\mathcal{D}_p$ correspond to the spot price of electricity in France. The data have been coupled with consumption and production data, since they are highly correlated: indeed, with very low or even zero marginal costs, renewable energy types are the first to be called upon. If renewable energy is sufficient to cover all demand, the price of spot electricity is close to 0€ per MWh.

## 2.2 Evaluation

The challenge is to forecast the net load in France. In fact, three components need to be taken into account in the assessment: total load, solar and wind production. Therefore, we can measure the performance of the models with the following measure:

$$\ell(y, \hat{y}) = \ell_{\mathbf{load}}(y, \hat{y}) + \ell_{\mathbf{solar}}(y, \hat{y}) + \ell_{\mathbf{wind}}(y, \hat{y}). \tag{1}$$

However, models may seek to optimize the loss directly in relation to the net load (without seeking to predict the various components), and so the measure can just be $\ell(y, \hat{y}) = \ell_{\mathbf{netload}}(y, \hat{y})$ – typically, MAPE or RMSE can be taken as a loss.

## 2.3 Models

In this section, we present the models developed as part of the challenge.Then, in order to make the most out of these models, participants are introduced to expert aggregation.

### 2.3.1 Baseline Models

**Persistence Model.** Persistence models are among the simplest forecasting models. The forecast at time $t$ is given by the value of the signal observed at $t - \Delta t$. where $\Delta t$ is the only parameter of the model. In this case, we choose a 1-day persistence.

**Gradient Boosting Models.** Gradient Boosting refers to a class of machine learning models that combine sequentially weak learners, typically decision trees, to create a strong final predictive model. The process involves iteratively fitting new models to the residual errors of the previous models, and this way to gradually improve the overall prediction. At each iteration, the new model is trained to minimize a loss function by adjusting its parameters in the direction that reduces the gradient of the loss. We use XGBoost and CatBoost (Chen and Guestrin, 2016, Prokhorenkova et al., 2019) as reference implementations for the Challenge.

**Generalized Additive Models.** Generalized Additive Models (GAMs) is a class of semi-parametric regression models that was first developed in (Hastie and Tibshirani, 1986) and (Hastie, 2017) and are now widely used in electricity consumption forecasting. Indeed, GAMs are interesting in practice, since their additive aspect makes them highly explainable, but this also means that the choice of variables must be meticulous. Consider a prediction model aiming to predict for each time $t$ a variable of interest $y_t$ using $(x_j)_{j=1\dots d}$ explanatory variables such that $y_t = X_t\beta_0 + \sum_{j=1}^{d} f_j(x_{t,j}) + \varepsilon_t$, where $\beta_0$ is the intercept, $X_t = [x_{t,1}, \dots, x_{t,d}]$ and $(\varepsilon_t)$ is i.i.d. random noise. Here we consider that each non-linear effect $f_j$ is decomposed on a spline basis $(B_{j,k})$ with coefficient $\mathbf{B_j} \in \mathbb{R}^{m_j}$ where $m_j$ is the chosen spline basis dimension, such that $f_j(x) = \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}(x)$. These coefficients are then estimated by minimizing the ridge-regression criterion ensuring the smoothness of the functions $f_j$ by controlling the second derivatives (Wood et al., 2016). Three GAM models were developed for the challenge, a model forecasting net load, a model forecasting load and total renewable production, and a model forecasting load, wind production and solar production. These models are respectively referred to as GAM-1, GAM-2 and GAM-3.

### 2.3.2 Aggregation of Experts

Several models have been developed in the literature, each with its own distinctive features, which may also complement each other. Expert aggregation is an ensemble technique that allows to benefit from the advantages of each model: we can combine them using robust online aggregation of experts, as developed in (Cesa-Bianchi and Lugosi, 2006). For each instant in the prediction, a weight is assigned to each expert according to its previous forecasts: the better the past forecasts, the greater the weight at time $t$. Let $x_{j,t}$ be the $j^{th}$ expert at time $t$ and $p_{j,t}$ its corresponding weight, then the expression of the predicted load at time $t$ is given by $\widehat{y}_t = \sum_{j=1}^{K} p_{j,t} x_{j,t}$, where $K$ is the number of experts in the mixture. One way to compute the weights is to use polynomially weighted averages with multiple learning rate (ML-Poly), an algorithm developed in (Gaillard et al., 2014). A key advantage lies in the upper bound of the algorithm's average error:

$$\textbf{Average error of the algorithm} \lesssim \textbf{Average error of the best combination of experts} + \sqrt{\frac{\textbf{Number of experts}}{\textbf{Number of days}}} \quad (2)$$

Over time, the algorithm's average performance will converge to match the performance of the best experts.

# 3 Results

## 3.1 Numerical Results

Numerical results at the national level are shown in Table 1.

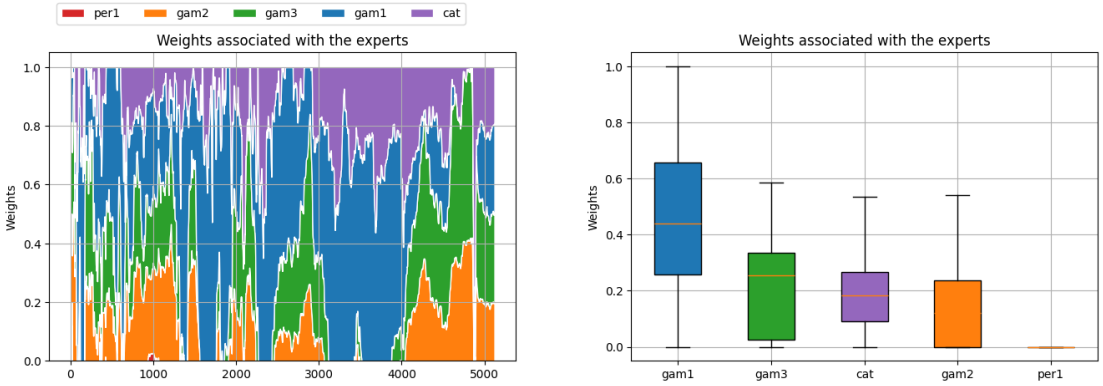
| Model | RMSE (MW) | MAPE (%) |
|---|---|---|
| GAM-1 | *1849* | *3.31* |
| GAM-2 | 1990 | 3.63 |
| GAM-3 | 1967 | 3.56 |
| CAT | 3341 | 5.99 |
| PER-1 | 5239 | 8.93 |
| Mixture | **1511** | **2.66** |

Table 1: Numerical performance in MAPE (%) and RMSE (MW) for $\mathcal{T}_{\text{fin}}^{\textbf{te}}$ at national level.

The hyperparameters of the baseline models above were optimized using the CMA-ES algorithm (Hansen et al., 2003).

## 3.2 Aggregation of Experts

Figure 4 shows the weights associated with each baseline expert.




(a) Weights associated to the experts. (b) Contibution of the experts.

Figure 4: Aggregation of the baseline models for $\mathcal{T}_{\text{fin}}^{\mathbf{te}}$.

# 4 Perspectives

These models do not take into account the spatial aspect of the data: data can be represented using graphs. We believe it is interesting to make use of the deep relationships that exist between the regions for forecasting as their features are strongly correlated, see Figure 5, and therefore to develop graph neural networks (GNNs) models for the Challenge which could be then be used as novel experts in an aggregation.

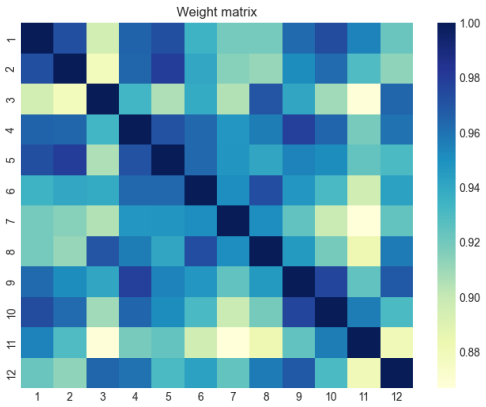



Figure 5: Correlation between the regions after a dimension reduction of the feature vector.

In graph theory, objects are represented by nodes, and the relationships between them are represented by edges. A graph $\mathcal{G}$ is a couple $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ a set of edges, i.e. $\mathcal{E} = \{(e_{ij}) = v_i v_j \mid v_i, v_j \in \mathcal{V}\}$. These graphs can be represented using adjacency matrices defined as $\mathbf{A} = (\mathbf{A}_{ij}) \in \mathbb{R}^{N \times N}$ such that $\mathbf{A}_{ij} = 1$ if and only if $e_{ij} \in \mathcal{E}$. Instead of a binary-weighted adjacency matrix, a more flexible weight matrix $\mathbf{W}$ with real-valued weights can be utilized. In the context of the Challenge, regions and the hidden links between them

can respectively be seen as nodes and edges. Both regression and classification tasks can be performed on graphs at different levels: node-level, edge-level and graph-level, see Figure 6.

- The **node-level** focuses on individual nodes within a graph. It involves analyzing the properties or attributes of each node. For example, in the context of electricity forecasting, a node-level task could be to predict the consumption for each region.
- The **edge-level** pertains to the analysis of the edges or connections between nodes in a graph. It involves examining the relationships, weights, or properties associated with each edge. For example, in the context of electricity forecasting, an edge-level task could be to quantify the relationships between the regions.
- The **graph-level** refers to the analysis of the entire graph structure as a whole. It involves examining global properties, overall connectivity, or emergent behaviors of the graph. For example, in the context of electricity forecasing, a graph-level task could be to predict the national consumption.

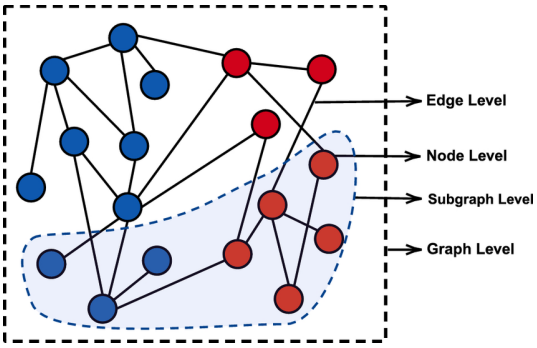



Figure 6: Various hierarchies of tasks in a graph representing edge, node, subgraph, and graph (Waikhom and Patgiri, 2022).

In the context of GNNs, message passing refers to the process of exchanging information between nodes, edges, and the global level of a graph (see Figure 7). Message passing is a fundamental operation in GNNs that enables nodes to gather and aggregate information from their neighbors, incorporate it into their own representations, and propagate it throughout the graph. Hence, a GNN corresponds to a set of layers that use the message-passing mechanism. Node representations are therefore updated as the graph is iterated through (in other words, at each layer traversed, representations are updated).

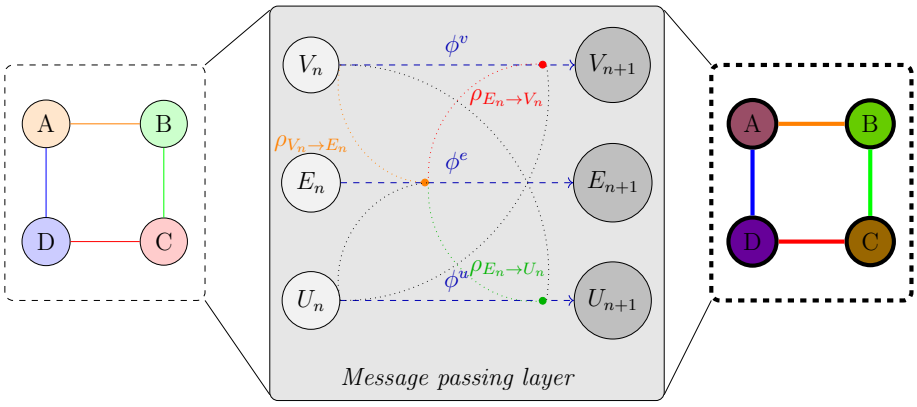



Figure 7: Example of a message passing layer in a GNN. $V_n$, $E_n$ and $U_n$ respectively refer to node, edge, and global level at stage $n$. $\phi$ are update functions and $\rho$ are propagation functions.

A message passing layer therefore enables a node to update its embedding by taking into account information from its neighbors: thus, by propagation, $d$ message passing layers enable a node to take into account information from its $d^{th}$-order neighbors. Using the notations of Figure 7, the message passing mechanism for each level can be written as follows:

- $V_{n+1} = \phi^v \left(V_n \ ; \ \rho_{E_n \to V_n}, \rho_{U_n \to V_n}\right)$
- $E_{n+1} = \phi^e \left(E_n \ ; \ \rho_{V_n \to E_n}, \rho_{U_n \to E_n}\right)$
- $U_{n+1} = \phi^u \left(U_n \ ; \ \rho_{V_n \to U_n}, \rho_{E_n \to U_n}\right)$

In a standard neural network, such as a feedforward neural network, the update process is typically a local operation. In contrast, message passing in GNNs is a more dynamic and relational process allowing nodes to capture both local and global graph structures and dependencies. This relational approach enables GNNs to capture complex graph patterns and dependencies that cannot be easily captured by standard neural networks, making them suitable for tasks involving graph-structured data. The message passing mechanism can also be seen as a special case of graph convolution. Convolutional operations, originally developed for regular grid-structured data such as images, enable the extraction of meaningful features by considering the local neighborhood of each element (LeCun et al., 1995). By extending convolution to graphs, we can capture and analyze the structural patterns and relationships present in the graph data, see (Daigavane et al., 2021) for a visual understanding. Graph convolution extracts localized features, where information from neighboring nodes is aggregated to compute features for each node allowing to capture the local connectivity and dependencies between nodes. It also helps to leverage the inherent structure and connectivity of the graph data, uncover hidden patterns (e.g. the relationship between the consumptions of two different regions), and make informed predictions (e.g. the consumption of a given region) based on the relationships between nodes.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, and for each node $v \in \mathcal{V}$, we associate a feature vector $X_v$ which corresponds to the explanatory variables of the corresponding node. Each node $v \in \mathcal{V}$ has an associated label $y_v$ and we want to learn a representation $h_v$ such that, for a GNN $f$, we have $f(h_v) = y_v$. To do this, we iteratively update the representations of a node by aggregating the representations of its neighbors in the graph. The representation of a node $v$ at iteration $k$, denoted $h_v^{(k)}$ can be expressed as

$$\begin{aligned} h_v^{(k)} &= \mathrm{UPDATE}^{(k)} \left( h_v^{(k-1)}; \mathrm{AGGREGATE}^{(k)} \left( h_v^{(k-1)}; \left\{ h_u^{(k-1)} \mid u \in \mathcal{N}_v \right\} \right) \right), \\ h_v^{(0)} &= X_v, \end{aligned} \tag{3}$$

where UPDATE usually involves combining the prior representations with the current one and a linear mapping. AGGREGATE is usually a combination of a pooling function (max, sum,...) with an activation function (ReLU, tanh,...). Two approaches are being studied for the Challenge: a basic graph convolutional network (GCN) model (Kipf and Welling, 2017) and the SAGE model (Hamilton et al., 2018). These models allow a gain of around 0.2% in the aggregation.

# References


Alasali, F., Nusair, K., Alhmoud, L., and Zarour, E. (2021). Impact of the covid-19 pandemic on electricity demand and load forecasting. *Sustainability*, 13(3):1435.

Antoniadis, A., Gaucher, S., and Goude, Y. (2022). Hierarchical transfer learning with applications for electricity load forecasting.

Brégère, M. and Huard, M. (2022). Online hierarchical forecasting for power consumption data. *International Journal of Forecasting*, 38(1):339–351.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Daigavane, A., Ravindran, B., and Aggarwal, G. (2021). Understanding convolutions on graphs. *Distill*, 6(9):e32.

de Vilmarest, J. and Goude, Y. (2021). State-space models win the ieee dataport competition on post-covid day-ahead electricity load forecasting.

Gaillard, P., Stoltz, G., and Van Erven, T. (2014). A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR.

Hamilton, W. L., Ying, R., and Leskovec, J. (2018). Inductive representation learning on large graphs.

Hansen, N., Müller, S. D., and Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11(1):1–18.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion), statistical science.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Obst, D., de Vilmarest, J., and Goude, Y. (2021). Adaptive methods for short-term electricity load forecasting during covid-19 lockdown in france. *IEEE Transactions on Power Systems*, 36(5):4754–4763.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2019). Catboost: unbiased boosting with categorical features.

Waikhom, L. and Patgiri, R. (2022). A survey of graph neural networks in various learning paradigms: methods, applications, and challenges. *Artificial Intelligence Review*.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

# Quantifying the Uncertainty of Electric Vehicle Charging with Probabilistic Load Forecasting

Yvenn Amara-Ouali [1] & Bachir Hamrouche [2] Guillaume Principato [3] & Yannig Goude [4]

[1] *University Paris-Saclay, LMO, France, yvenn.amara-ouali@universite-paris-saclay.fr*
[2] *EDF R&D, OSIRIS, Palaiseau, France, bachir.hamrouche@edf.fr*
[3] *University Paris-Saclay, LMO, France, guillaume.principato@universite-paris-saclay.fr*
[4] *EDF R&D, OSIRIS, Palaiseau, France, France et yannig.goude@edf.fr*

**Résumé.** Cet article explore les moyens de quantifier l'incertitude associée à l'utilisation croissante des véhicules électriques (VE) dans la gestion des réseaux électriques. En mettant l'accent sur les solutions de prévision de la charge, l'étude étend un premier benchmark effectué sur la prévision de la charge à l'horizon J+1 pour inclure des algorithmes de prévision probabilistes. Deux approches sont envisagées : une approche directe qui fournit des prévisions de quantiles à l'aide de modèles GAMlss, et une approche *bottom-up* qui prédit les sessions de charge individuelles avant de reconstruire la courbe de charge globale et de calculer les quantiles empiriques. Les méthodes proposées sont évaluées à l'aide de mesures telles que Pinball Loss et l'erreur quadratique, démontrant des performances comparables entre l'approche directe et l'approche ascendante. Les résultats, basés sur des données réelles de sessions de charge à Palo Alto, suggèrent les avantages potentiels de l'approche ascendante pour les quantiles élevés.

**Mots-clés.** Prévisions Quantiles, Monte-Carlo, Approche *Bottom-Up*, Modèles Additifs Généralisés, Modèles de Mélange, Prédiction Conforme, *Smart Grids*

**Abstract.** This paper explores ways of quantifying the uncertainty associated with the increasing use of electric vehicles (EVs) in the management of electricity networks. With a focus on load forecasting solutions, the study extends a benchmark on day-ahead load forecasting to include probabilistic forecasting algorithms. Two approaches are considered: a direct approach that provides quantile forecasts using GAMlss models, and a bottom-up approach that predicts individual charging session characteristics before reconstructing the load curve and calculating empirical quantiles. The proposed methods are evaluated using metrics such as Pinball Loss and RMSE, demonstrating comparable performance between the direct and bottom-up approaches. The results, based on real data from Palo Alto charging sessions, suggest potential advantages of the bottom-up approach for high quantiles.

**Keywords.** Quantile forecasts, Monte-Carlo Simulation, Bottom-up approach, Generalised Additive Models, Gaussian Mixtures, Conformal Prediction, Smart Grids

# 1 Industrial Context

A key lever for reducing greenhouse gas emissions in the transport sector is the large-scale deployment of electrical vehicles (EVs). This has led to many governments implementing strong pro-EV policies, resulting in an increase in the number of electrical vehicles in global markets. The arrival of these vehicles creates challenges in the management of the electrical network while also bringing opportunities in terms of grid flexibility. Indeed, these vehicles will become important assets for managing electricity demand whose charging can be automatically postponed when constraints are high or even used as storing batteries to reinject power when demand is high. All these operations of load optimisation are referred to as smart charging. One of the key elements of an efficient smart charging solution is a strong understanding of charging behaviours which requires the development of efficient forecasting algorithms in order to predict them.

# 2 Related Work

This work builds on a previous set of papers [1] [2] which focused on benchmarking day-ahead load and occupancy forecasting solutions, mainly using algorithms that return point estimates corresponding to the mean of the distribution of interest. The previous benchmark examined two sets of methods: direct approaches that would predict the aggregate load curve at the station level, and bottom-up approaches that would model the set of individual behaviours before aggregating them to obtain the predicted curve at the station level. The bottom-up approaches, although more complex to estimate, offer more flexibility for the deployment of smart charging solutions. In this work, we propose to extend this benchmark to probabilistic forecasting algorithms by exploring probabilistic variations. The use of probabilistic forecasts is becoming increasingly important for the efficient operation of electricity systems, as highlighted in the last two Global Energy Forecasting Competitions [3][4]. Recently, several approaches for estimating probabilistic forecasts related to energy demand have been proposed in [5] and [6]. The need for probabilistic forecasts is particularly important in the management of electric vehicles, as the optimisation of charging loads often requires a good quantification of the certainty around the forecasts in order to manage the best and worst case scenarios. [7] proposes an approach to quantify the uncertainty of parking duration forecasts in EV management.

# 3 Methods

Two approaches have been used to address the probabilistic forecasting task. A direct approach which provides quantile forecasts with a GAMlss (see Section 3.1) model trained directly on the load curve. A bottom-up approach (see Section 3.2) which predicts individual charging session characteristics to then reconstruct the load curve. With both approaches, 9 quantile forecasts are provided from 0.1 to 0.9 with 0.1 increments.

### 3.1 Direct Approach

Generalised Additive Models for location scale and shape (GAMlss) are an extension of GAM [10] which enables the fine modelling multiple parameters of a single distribution. In this study, GAMlss are used to model both the mean and variance of the load at charging points over time.

## 4 Bottom-Up Approach

Bottom-up approaches predict the characteristics of individual charging sessions occurring over time. In particular, three variables are required to reconstruct the load curve of an ensemble of charging stations in an uncontrolled charging environment: $(a_i, d_i, e_i)$ the arrival time, charging duration and energy demand of a charging sessions $i$. These three variables can be modelled using various statistical techniques. It was shown in [2] and [3] that mixture models are an adequate choice of method to represent individual charging sessions. Assuming we can predict the number of charging sessions $N$ occurring each day with a time series model, we can sample from the mixture model distribution N times to obtain a prediction for a particular day. In this work, a SARIMA model is used as the predictor for the number of daily charging sessions. A Monte-Carlo simulation is executed to obtain empirical quantiles of the SARIMA model. For each of these 9 forecasted quantiles, another Monte-Carlo simulation is led on the mixture models to reconstruct a total of 10000 load curves from which empirical quantiles are recovered for each instant.

### 4.1 Metrics

Three types of metrics have been used to evaluate quantile forecast performances. First the pinball loss defined as follows:

$$L_{\tau}^{\text{pinball}}(y_{\text{obs}}, (\hat{y}_{\tau})) = \frac{1}{N} \sum_{i=1}^{N} \rho_{\tau}(y_{\text{obs}}^{i} - (\hat{y}_{\tau}^{i})) \tag{1}$$

$$\rho_{\tau}(u) = \begin{cases} \tau \cdot u, & \text{if } u < 0 \\ (1-\tau) \cdot u, & \text{if } u \geq 0 \end{cases} \tag{2}$$

It penalises the model for deviations between the true target values $y_{obs}$ and the predicted quantiles $(y_{\tau})$. The degree of penalty depends on the chosen quantile level $\tau$. Another metric which can be used for assessing the accuracy of quantile forecasts can be defined as follows:

Equation (3) provides an estimate of the probability of the observed value falling below the predicted quantile. Essentially, if $L_{\tau}^{emp}(y_{obs}, (y_{\tau})) = \tau$ the quantile forecast is optimal on the testing sample.

$$L_\tau^{\text{emp}}(y_{\text{obs}}, (\hat{y}_\tau)) = \frac{1}{N}\sum_{i=1}^{N} I(y_{\text{obs}}^i \leq (\hat{y}_\tau^i)) \tag{3}$$

A slightly modified version of the metric defined in equation (3) can be written as follows using block-bootstrap sampling:

$$L_\tau^{\text{rmse}}(y_{\text{obs}}, (\hat{y}_\tau)) = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\tau - \frac{1}{N}\sum_{i=1}^{N} I(y_{\text{obs}}^i \leq (\hat{y}_\tau^i))\right)^2} \tag{4}$$

With B=1000 the number of bootstrap samples of size N.

# 5 Results and Discussion

Experiments were led on the city of Palo Alto (California, USA) dataset that gathers real data from charging sessions occurring. This data has been explored in [1]. Figure 1 shows the average daily quantile forecasts for both approaches. It seems that both approaches capture the general shape of the observed curve in black with a peak demand in the middle of the day and another one in the evening. Both models seem to slightly underestimate high quantiles. This is confirmed by Figure 2 where it can be observed that $L_\tau^{emp} \leq \tau$ for all quantiles except 0.1. Figure 2 also shows that both the direct approach and the bottom-up approaches yield similar performances also confirmed by Table 1. It also indicated that the bottom-up approach could be more performant for high quantiles and particularly quantile 0.9 which is more performant for the bottom-up approach across all metrics considered.

Table 1: Pinball Losses ($L_\tau^{\text{Pinball}}$) and RMSE ($L_\tau^{\text{rmse}}$) metrics for each quantile level

| Quantile | Pinball Loss | | RMSE | |
|---|---|---|---|---|
| | GAMlss | Bottom-Up | GAMlss | Bottom-Up |
| 0.1 | **2.59** | 2.65 | **0.0227** | 0.0356 |
| 0.2 | **4.25** | 4.36 | **0.0165** | 0.0198 |
| 0.3 | **5.36** | 5.53 | **0.0401** | 0.0503 |
| 0.4 | **6.02** | 6.27 | **0.0628** | 0.0764 |
| 0.5 | **6.31** | 6.60 | **0.0758** | 0.101 |
| 0.6 | **6.22** | 6.48 | **0.0892** | 0.115 |
| 0.7 | **5.72** | 5.92 | **0.110** | 0.122 |
| 0.8 | **4.74** | 4.82 | 0.113 | **0.107** |
| 0.9 | 3.13 | **3.05** | 0.106 | **0.0766** |

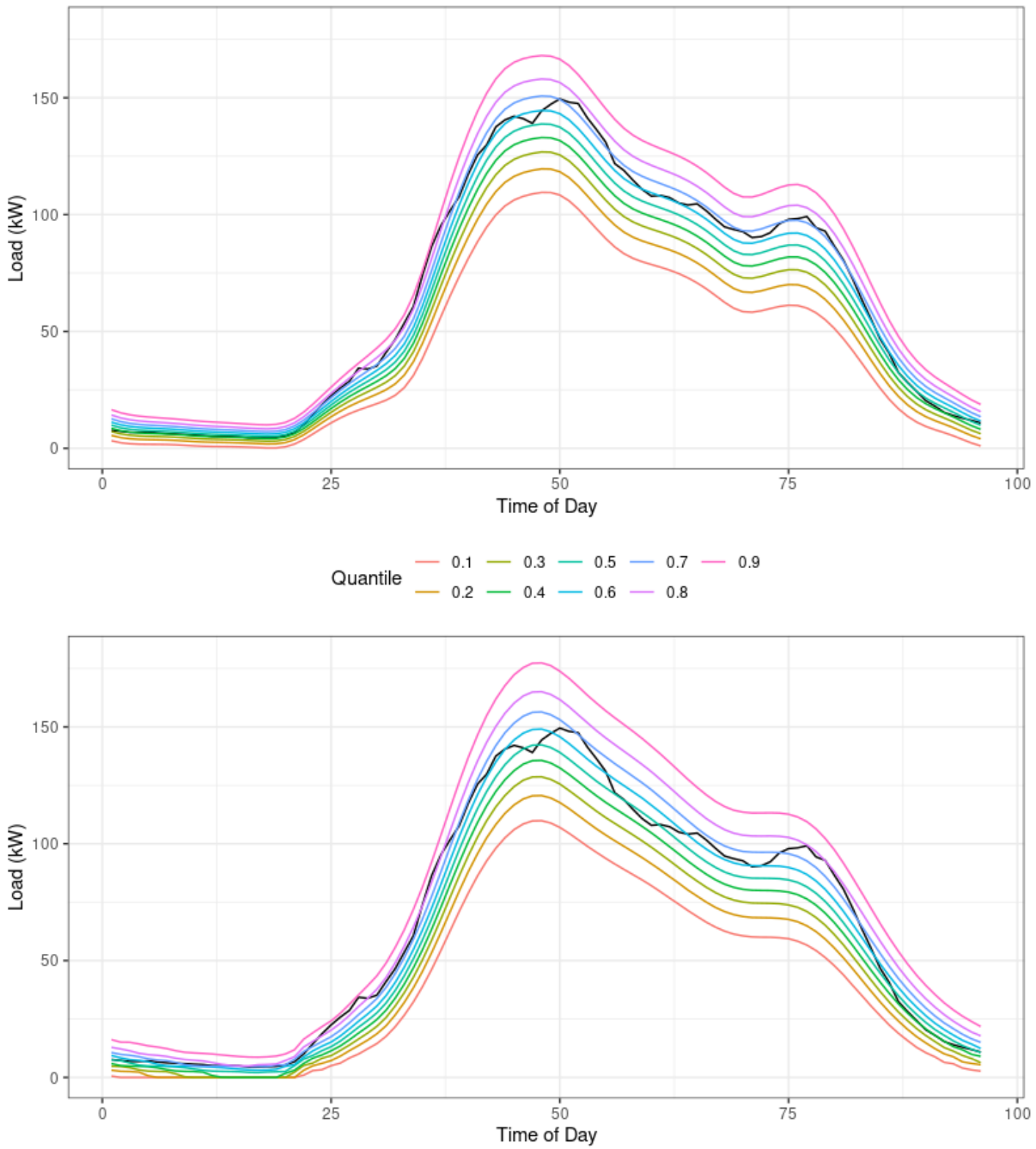



Figure 1: At the top are the GAMlss averaged quantile forecasts for each quantile levels and at the bottom the same chart for the bottom-up approach. In black is the average daily load curve over the test period.

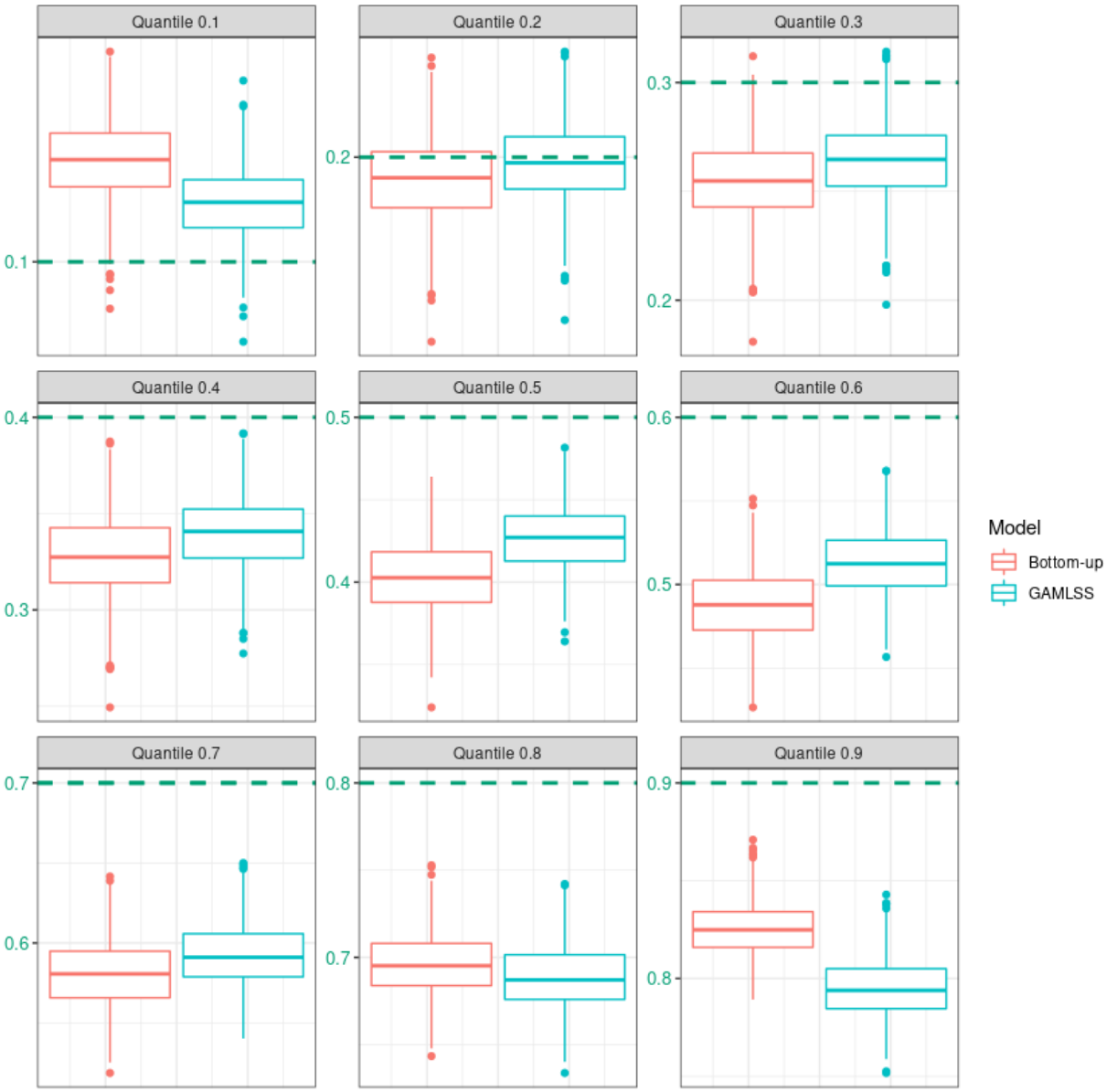



Figure 2: Boxplots of the block-bootstrap performances calculated with $L_\tau^{emp}$ defined in equation (3)

## Acknowledgements

The results presented in this abstract will be further detailed and complemented with conformal predictions after review and upon acceptance.

## References

[1] Amara-Ouali, Y., Goude, Y., Massart, P., Poggi, J. M., & Yan, H. (2021). A review of electric vehicle load open data and models. *Energies*, 14(8), 2233.

[2] Amara-Ouali, Y., Goude, Y., Hamrouche, B., & Bishara, M. (2022, June). A benchmark of electric vehicle load and occupancy models for day-ahead forecasting on open charging session data. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems* (pp. 193-207).

[3] Hong, T., et al. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of forecasting*, 32(3), 896-913.

[4] Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 35(4), 1389-1399.

[5] de Vilmarest, J., et al. (2023). Adaptive Probabilistic Forecasting of Electricity (Net-) Load. arXiv preprint arXiv:2301.10090.

[6] Capezza, C., et al. (2021). Additive stacking for disaggregate electricity demand forecasting. *The Annals of Applied Statistics*, 15(2), 727-746.

[7] Phipps, K., et al. (2023). Customized Uncertainty Quantification of Parking Duration Predictions for EV Smart Charging.

[8] Flammini, M. G., et al. (2019). Statistical characterisation of the real transaction data gathered from electric vehicle charging stations. *Electric Power Systems Research*, 166, 136-150.

[9] Barman, P., et al. (2023). Renewable energy integration with electric vehicle technology: A review of the existing smart charging approaches. *Renewable and Sustainable Energy Reviews*, 183, 113518.

[10] Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3), 297.

# Une politique d'inspections et de remplacements pour un modèle de dégradation avec effets de maintenance partiels

Margaux Leroy [1] Christophe Bérenguer [2] Laurent Doyen [3]

[1] *Université Grenoble Alpes, CNRS, Grenoble INP, Gipsa-LAB, LJK*
*margaux.leroy@univ-grenoble-alpes.fr*
[2] *Université Grenoble Alpes, CNRS, Grenoble INP, Gipsa-LAB,*
*christophe.berenguer@grenoble-inp.fr*
[3] *Université Grenoble Alpes, CNRS, Grenoble INP, LJK,*
*laurent.doyen@univ-grenoble-alpes.fr*

**Résumé.** Ce papier se concentre sur l'optimisation de la modélisation de coût d'une politique de maintenance, initialement basée sur un modèle de dégradation construit à partir de deux processus de Wiener dépendants avec des effets de maintenance partiels. La politique de maintenance proposée implique des inspections, des réparations (ou maintenances imparfaites) et des remplacements dits "AGAN" (As-Good-As-New). Le temps d'inter-inspection et le seuil de remplacement préventif sont optimisés en minimisant le coût de maintenance asymptotique par unité de temps. Ce coût est déterminé en utilisant les propriétés de renouvellement Markovien du processus de dégradation sur le système maintenu. Deux méthodes sont envisagées pour calculer et minimiser le coût asymptotique par unité de temps et obtenir la politique de maintenance optimale.

**Mots-clés.** Processus stochastique, processus de Wiener, modèle de dégradation, maintenances imparfaites, politique de maintenance, optimisation de coût

**Abstract.** This paper focuses on optimizing the cost modeling of a maintenance policy, initially based on a degradation model built from two dependent Wiener processes with partial maintenance effects. The proposed maintenance policy involves inspections, imperfect repairs, and so-called "AGAN" (As-Good-As-New) replacements. The inter-inspection time and the preventive replacement threshold are optimized by minimizing the asymptotic maintenance cost per time unit. This cost is determined using the Markov renewal properties of the degradation process on the maintained system. Two methods are considered to assess and minimize the asymptotic cost per time unit and derive the optimal maintenance policy.

**Keywords.** Stochastic process, degradation model, imperfect maintenance, maintenance policy, cost optimization

# 1 Modélisation de la dégradation en présence de maintenances

Le modèle de dégradation étudié est construit à partir de deux processus de Wiener dépendants $X^U = \{X^U(t)\}_{t\geq 0}$ et $X^M = \{X^M(t)\}_{t\geq 0}$. Cette dégradation est altérée par divers types d'actions de maintenance : des inspections, des réparations imparfaites et des remplacements dits "AGAN" (As-Good-As-New), en ce sens qu'après chaque remplacement le système est comme neuf et la dégradation retombe à son état initial. Entre deux maintenances, la dégradation évolue en tant que $\{X^S(t)\}_{t\geq 0}$, somme de ces deux processus, telle que $X^S(t) = X^U(t) + X^M(t)$. Aux instants de réparation imparfaites, seul le processus $X^M$ est affecté par l'effet de la maintenance (Leroy et al. 2024). Cet effet partiel est immédiat et de type $ARD_1$ (Mercier and Castro, 2019) : le niveau de dégradation est réduit proportionnellement à la dégradation accumulée sur le processus maintenu depuis la dernière maintenance.

La dégradation du système est inspectée périodiquement avant chaque réparation. Les instants de maintenance sont notés $\{\tau_j\}_{j\geq 0}$ et $\tau_0 = 0$. La période entre deux inspections est notée $\widetilde{\tau}$, telle que $\tau_j = j\widetilde{\tau}$. $Y(t)$ exprime le niveau de dégradation à l'instant $t$ et $Y(\tau_j^-)$, resp. $Y(\tau_j^+)$, représente le niveau de dégradation juste avant, resp. juste après, la $j^{ieme}$ maintenance. A la suite d'une réparation imparfaite, le niveau de dégradation équivaut à : $Y(\tau_j^+) = X^S(\tau_j) - \rho\ X^M(\tau_{j-1})$, $\forall\ j \geq 1$, où $\rho \in [0, 1]$ représente le paramètre d'efficacité de la maintenance.

# 2 Hypothèses de maintenance

Le niveau de dégradation inspecté juste avant chaque maintenance, $Y(\tau_j^-)$, détermine la nature de l'intervention : réparation imparfaite ou remplacement, tous deux réalisés instantanément. $M$ représente le seuil de remplacement préventif et $L$ le seuil de remplacement correctif. Comme illustré sur la Figure 1, pour chaque intervalle $[\tau_{j-1}, \tau_j[$, la décision du type de maintenance s'appuie sur les hypothèses suivantes :

- S'il existe $t \in\ [\tau_{j-1}, \tau_j[$ tel que $Y(t) > L$, alors une panne du système survient au temps $t$ lorsque la dégradation dépasse pour la première fois le seuil $L$, entraînant une indisponibilité temporaire du système. Lors du prochain instant de maintenance, l'ensemble du système est soumis à un remplacement correctif, le niveau de dégradation retombe donc à zéro: $Y\left(\tau_j^+\right) = 0$. Si la dégradation ne dépasse pas le seuil $L$ entre deux instants d'inspection, alors soit un remplacement préventif soit une réparation imparfaite est effectué sur le système.
- Si $Y\left(\tau_j^-\right) < M$, une réparation imparfaite est réalisée.
- Si $M \leq Y(\tau_j{}^-) \leq L$, un remplacement préventif est effectué sur l'intégralité du système, initiant un nouveau cycle de vie tel que $Y(\tau_j{}^+) = 0$.

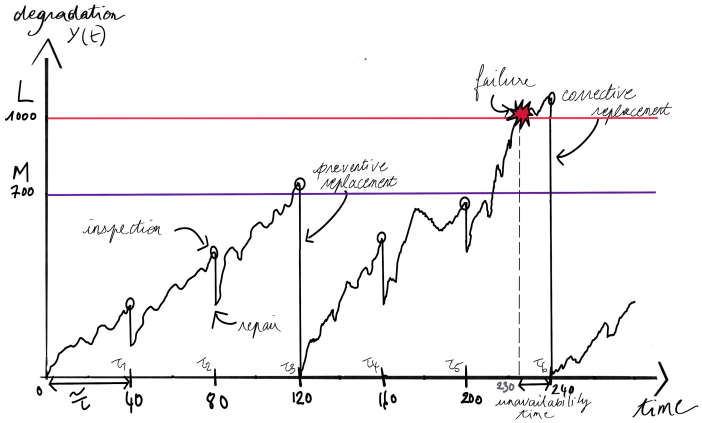



Figure 1: Schéma représentant l'évolution de la dégradation $Y(t)$ d'un système soumis à différent types de maintenance

Les variables de décision étudiées pour la politique de maintenance envisagée sont la période d'inter-inspection ($\widetilde{\tau}$) et le seuil de remplacement préventif $M$.

# 3 Optimisation de la politique de maintenance $(\widetilde{\tau}, M)$

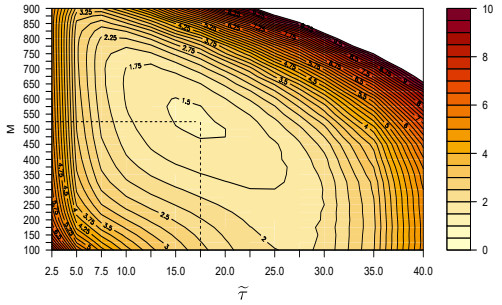



Figure 2: Optimisation de la politique de maintenance $(\widetilde{\tau}, M)$ en minimisant le coût asymptotique par unité de temps

Chaque opération de maintenance engendre un coût spécifique. La politique de maintenance optimale $(\widetilde{\tau}^*, M^*)$ est déterminée en minimisant le coût asymptotique par unité de temps. La difficulté d'exprimer analytiquement le coût asymptotique par unité de temps peut être appréhendée en considérant le processus de dégradation $\{Y(t)\}_{t\geq 0}$ comme un processus de renouvellement Markovien (Cocozza-Thivent 2000). En effet, après chaque remplacement, la dégradation redémarre de zéro et le processus qui suit ne dépend pas des évènements antérieurs à ce remplacement. Ainsi, le théorème de renouvellement peut s'appliquer et optimiser le coût asymptotique par unité de temps revient à optimiser l'espérance du coût sur un cycle (entre deux remplacements) divisé par l'espérance d'un cycle.

Une approche alternative consiste à optimiser le coût asymptotique par unité de temps entre deux inspections (Cocozza-Thivent, 2000, Grall and al. 2002, Corset and al. 2022). Sur $[0, \widetilde{\tau}]$, $Y(\widetilde{\tau}^-)$, le niveau de dégradation en $\widetilde{\tau}$ juste avant que la maintenance ne soit réalisée, ne dépend que de $Y_0$, le niveau de dégradation à l'état présent, dont la loi est stationnaire. Le processus de dégradation $\{Y(t)\}_{t\geq 0}$ peut alors être décrit comme un processus de Markov semi-régénératif entre deux inspections. Ainsi, une mesure stationnaire $\pi$ peut être appliquée à la chaîne de Markov $(Y_n)_{n\in\mathbb{N}}$ sur l'espace d'état continu $\mathbb{R}$, où $(Y_n)_{n\in\mathbb{N}}$ représente les niveaux de dégradation juste après chaque action de maintenance (réparation imparfaite ou remplacement), tel que $(Y_n)_{n\in\mathbb{N}} = Y(\tau_j^+)_{j\geq 1}$ . D'après (Cocozza-Thivent, 2000), le coût asymptotique par unité de temps est alors égal à l'espérance du coût calculé sur $[0, \widetilde{\tau}]$ sous la mesure stationnaire, divisé par $\widetilde{\tau}$. On peut ainsi de nouveau évaluer et minimiser ce coût asymptotique pour obtenir une politique optimale de maintenance en fonction des variables de décision $(\widetilde{\tau}, M)$.

Ces deux approches sont utilisées dans la détermination de la politique de maintenance optimale $(\widetilde{\tau}^*, M^*)$ et leur efficacité est comparée. Par ailleurs, l'influence des paramètres du modèle et des coefficients de coût sur cette politique optimale est également étudiée.

# Bibliographie


Cocozza-Thivent C., 2000, Convergence de fonctionnelles de processus semi-régénératifs, Prépublications de l'Université Marne la Vallée, no. 2/2000.

Corset F., Fouladirad M., Paroissin C., 2022, Imperfect condition-based maintenance for a gamma degradation process in presence of unknown parameters, Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability.

Grall A, bérenguer C., Dieulle L., 2002, A condition-based maintenance policy for stochastically deteriorating systems, Reliability Engineering & System Safety 76, no 2.

Leroy M., Bérenguer C., Doyen L., Gaudoin O., 2024 (in press), Modelling and inference for a degradation process with partial maintenance effects.

Mercier S., Castro I.T., 2019, Stochastic comparison of imperfect maintenance models for a Gamma deteriorating system. European Journal of Operational Research , no. 1, 237-248.

# Séries temporelles 2

# Les erreurs-types dans les modèles non linéaires tels que les modèles de séries temporelles

Guy Mélard [1]

[1] *Université libre de Bruxelles, Belgique, gmelard@ulb.ac.be*

**Résumé.** L'estimation des paramètres de modèles statistiques non linéaires est généralement effectuée par une méthode de moindres carrés non linéaires ou de maximum de vraisemblance en employant de l'optimisation numérique. Les erreurs-types sont alors souvent déduites en employant des dérivées numériques. Une étude empirique dans le contexte de modélisation de séries temporelles au moyen de plusieurs logiciels statistiques révèle que ces erreurs-types ne sont pas très précises, correctes parfois à seulement 2 ou 3 chiffres significatifs. Une investigation complémentaire sur des modèles encore plus simples détermine la raison principale de ce manque de précision. On fournit plusieurs suggestions aux développeurs de logiciels statistiques dans le but d'améliorer leurs produits. Cette communication s'inscrit dans une suite de travaux de quelques auteurs qui mettent en gardent les utilisateurs trop confiants dans la pertinence de leurs résultats.

**Mots-clés.** Optimisation non linéaire, matrice d'information de Fisher, dérivées numériques, modélisation de séries temporelles.

**Abstract.** Estimation of the parameters in non-linear statistical models is usually performed by non-linear least-squares or maximum likelihood using numerical optimization. Then, standard errors are often derived by using numerical derivatives. An empirical study of time series modelling through several statistical packages reveals that these standard errors are not very accurate, sometimes with only 2 or 3 correct digits. A further investigation on much simpler models determines the main reason for such a lack of accuracy. Some suggestions are given to developers of statistical software in order to improve their products. The present communication is inline with a sequence of works from different authors who warn users that are too confident in the relevance of their results.

**Keywords.** Non-linear optimization, Fisher information matrix, numerical derivatives, time series modelling.

## 1 Introduction

Cette communication s'inscrit dans une suite de travaux de quelques auteurs qui mettent en gardent les utilisateurs trop confiants dans la pertinence de leurs résultats : McCullough (1998, 1999, 2004), McCullough et Renfro (2000), Yalta et Jenal (2009), et Hill et al. (2024), notamment.

Newbold et al. (1994) ont montré que différents logiciels conduisent, pour un modèle de séries temporelles spécifié, à des estimations et des erreurs-types différentes. Deux explications sont liées à la méthode d'estimation (moindres carrés conditionnels, pseudo-maximum de vraisemblance gaussien) et aux options spécifiques à l'optimisation. Les algorithmes utilisés pour les calculs sont aussi de première importance.

Nous avons d'abord voulu étendre cette vielle étude empirique à d'autres séries avec des modèles plus variés. On pourrait croire que ces différences ont disparu avec le temps mais ce n'est pas le cas, ou que des packages très employés comme R fournissent de meilleurs résultats mais c'est faux.

Notons que nous ne considérons ici que les erreurs-types déduites lors de l'estimation des paramètres d'un modèle. En supposant la normalité du processus générateur, il est aussi possible d'évaluer la matrice d'information de Fisher en tout point et, après inversion de la matrice en l'optimum, la matrice de variance-covariance des estimateurs des paramètres de modèles ARMA (et même VARMAX, c'est-à-dire ARMA multivariés et avec variables explicatives), voir Mélard et Klein (2023) et Klein et Mélard (2023). Nous ne considérons pas cette approche ici.

Partant d'une série $\{y_t, t = 1, ..., n\}$ de longueur $n$, et un modèle dépendant d'un vecteur de paramètres $\beta$, de vraie valeur inconnue $\beta^0$, on considère la log-vraisemblance $\ell_n(\beta)$ dont la maximisation conduit à un estimateur $\widehat{\beta}_n$. La matrice de variance-covariance de cet estimateur peut s'obtenir par une des deux espérances mathématiques suivantes

$$I(\beta^0) = \frac{1}{n} E_{\beta_0}\left(\frac{\partial \ell_n(\beta)}{\partial \beta^\top}\frac{\partial \ell_n(\beta)}{\partial \beta}\right) \quad \text{ou} \quad J(\beta^0) = -\frac{1}{n} E_{\beta_0}\left(\frac{\partial^2 \ell_n(\beta)}{\partial \beta^\top \partial \beta}\right), \tag{1}$$

où $\top$ indique la transposition. Mais $\beta^0$ est inconnu, les espérances sont pratiquement impossibles à évaluer et la forme analytique de $\ell_n(\beta)$ est trop complexe, de sorte qu'on est obligé d'employer une évaluation numérique en $\beta = \widehat{\beta}_n$. On n'a pas $I(\widehat{\beta}_n) = J(\widehat{\beta}_n)$ et il y a plusieurs manières d'évaluer le produit extérieur des gradients pour $I$ ou la hessienne pour $J$.

# 2 Description de la partie expérimentale

Nous sommes partis d'un ensemble de séries temporelles déjà considérées et de modèles ARIMA proposés dans la littérature, voir Mélard (1985). Nous avons estimé les paramètres de ces modèles et déterminé les erreurs-types au moyen d'un certain nombre de logiciels, parmi lesquels les packages stats (fonction arima) et forecast (fonction Arima) de R, SPSS, SAS, Stata et quelques autres. Notons que tous ces programmes utilisent le fonction de vraisemblance gaussienne exacte, contrairement à Newbold et al. (1994), donc les résultats sont plus proches. Nous avons choisi comme référence un programme personnel basé sur Mélard (1984), écrit spécialement et compilé en quadruple précision donc traitant des nombres avec plus 26 décimales. Le nombre de décimales correctes est calculées par la formule proposée par McCullough (1998), $-\log_{10}(|q - c|/|c|)$, où le résultat obtenu $q$ est comparé à la vraie valeur $c$, sauf si $c = 0$, où on prend $-\log_{10}(|q|)$

Nous avons alors examiné les nombres de chiffres corrects, synthétisés sur les différents paramètres, pour les estimations et pour les erreurs-types. Il n'est pas possible de dćrire ici les 27 séries de longueur $n$ entre 60 et 369, les spécification ARIMA utilisées avec entre 1 et 5 paramètres, les logiciels comparés avec les options appropriées choisies, ni de donner les résultats détaillés. En moyenne sur les séries, on observe que le nombre de chiffres corrects des erreurs-types 1 (pour les fonctions arima et Arima, respectivement dans les packages stats et forecast de R, et pour Stata), 3 (SAS), 4 (SPSS) sont bien inférieurs à ceux pour les estimations 3 (SPSS), 4, (la fonction arima du package stats de R et SAS), 5 (la fonction Arima du package forecast de R), et 6 (Stata). Notons que, contrairement aux autres, Stata inclut la variance des erreurs $\sigma^2$ parmi les paramètres d'intérêt ce qui a nécessité des ajustements. Dans le tableau 1, ils sont indiqués par la mention "ajust. pour variance".

# 3 Approche computationnelle

Comme nous l'indiquons dans Mélard (2024), les auteurs de logiciels ne documentent pas suffisamment les algorithmes employés, par exemple la méthode de calcul de la fonction objectif (parfois, mais pas toujours, la quasi-vraisemblance gaussienne exacte), l'algorithme d'optimisation, les critères d'arrêt, etc. On peut supposer que l'obtention des erreurs-types se fait par recours à des différences divisées, qui entraînent une perte de précision substantielle. Commme la plupart des logiciels (sauf Stata), on emploie la vraisemblance concentrée par rapport à la variance $\sigma^2$ des erreurs, donc on est ramené á minimiser une somme de carrés $S(\beta)$. Le minimum est alors divisé par $\widehat{\sigma}_n^2$, une estimation non biaisée de $\sigma^2$.

Au lieu de $S$, on peut aussi se baser sur les résidus $e_t$ dont $S$ est la somme des carrés. Il n'est pas possible de détailler ici les différentes approximations de $I$ et de $J$ qui ont été traitées. Pour examiner plus précisément ces aspects, nous avons traité des modèles non linéaires, puis des modèles linéaires pour terminer par l'estimation de la moyenne, c'est-à-dire d'une constante $c$ dans un modèle de série temporelle $y_t = c + e_t$. En effet, déjà dans ce cas-là, on va voir que les logiciels donnent des estimations peu correctes des erreurs-types et expliquer pourquoi. L'estimation optimale de $c$ est $\bar{y}$, la moyenne des $y_t$, donc $S(\bar{y})$ est $n$ fois la variance des $y_t$.

Nous avons expérimenté avec notre propre programme et développé plusieurs approches, présentées succinctement dans le tableau 1 et détaillées dans Mélard (2024), de manière à améliorer assez fortement la précision des résultats. A titre d'exemple, nous avons pris une des séries de longueur $n = 106$. Des expressions de $n\sigma^2$ fois l'information de Fisher $I$ (basée sur le produit extérieur des gradients ou OPG) ou $J$ (basée sur le hessien) sont proches de $n$. On devrait retrouver que l'erreur-type sur la moyenne vaut $1/\sqrt{n}$ fois l'écart-type estimé $\hat{\sigma}$. Ces approches sont comparées dans le tableau 1 qui montre qu'il est possible de passer de 2 à 9 chiffres corrects pour les erreurs-types. Nous avons ajouté les résultats obtenus par quelques logiciels déjà considérés au paragraphe 2. C'est évidemment avec cette variante dans notre programme en quadruple précision que les résultats du paragraphe 2 ont été obtenus. L'amélioration inspirée par Goldberg (1991) dont il est question dans le tableau 1 consiste simplement à exploiter l'identité $(a^2 - b^2) = (a+b)(a-b)$ qui se révèle très utile pour réduire les pertes de précision lors de différences entre deux valeurs de $e_t^2$ en des valeurs proches des

| Nom : méthode | Valeur | Précision | Err.-type | # décimales |
|---|---|---|---|---|
| $J_1$: $J$ avec $1^e$ dérivée des $e_t$ | 106,00000000 | 0 | 0,352742733 | |
| $I$ | 106,00000010 | $10^{-7}$ | 0,352742733 | 9,3 |
| $J_2$ (ancien): $J$ avec $2^{de}$ dérivée de $S$ | 110,61729310 | 4,6 | 0,345302319 | 1,7 |
| $J_2$ (nouveau) : idem, (somme avec DOT_PRODUCT) | 106,00042515 | $4\ 10^{-4}$ | 0,352742026 | 5,7 |
| $J_2$ (amélioré): $2^{de}$ dérivée des $e_t^2$, (somme avec SUM) | 106,00000212 | $2\ 10^{-6}$ | 0,352742729 | 8,0 |
| $J_2$ (amélioré bis): idem, (somme avec TwoSum, d'Ogita et al. (2005)) | 106,00000410 | $4\ 10^{-6}$ | 0,352742726 | 7,7 |
| $J_2$ (final): idem avec Goldberg (1991) (somme avec SUM) | 106,00000015 | $2\ 10^{-7}$ | 0,352742733 | 9,2 |
| R forecast/Arima | | | 0,351074987 | 2,3 |
| SAS | | | 0,352742733 | 10,4 |
| Stata OIM ajust. pour variance | | | 0,35107491 | 2,3 |
| Stata OPG ajust. pour variance | | | 0,35107491 | 2,3 |

Tableau 1: Résultats de $n\widehat{\sigma}^2 I$ ou de $n\widehat{\sigma}^2 J$ pour l'estimation d'une constante sur une série avec $n = 106$. SUM et DOT_PRODUCT calculent respectivement une somme et un produit scalaire avec une précision étendue au sens de la norme IEEE 754

paramètres.

# 4 Conclusions

Il est assez surprenant que des erreurs-types soient aussi peu précises dans les logiciels commerciaux ou même dans les logiciels libres utilisés partout.

On a aussi remarqué que les packages de R n'ont pas pu traiter tous les modèles, et que pour SAS et SPSS les options lors de l'optimisation doivent impérativement être modifiées (ce qui a été fait ici) sous peine d'avoir des résultats moins corrects.

On peut raisonnablement penser que nos constatations s'étendent à d'autres modèles statistiques et d'autres logiciels où l'optimisation numérique est employée, tout au moins quand la matrice de variance-covariance n'est pas obtenue en employant des dérivées analytiques de la log-vraisemblance.

# Bibliographie

Goldberg, D. (1991), What every computer scientist should know about floating-point arithmetic, *ACM Computing Surveys* **23**(1), 5-48.

Hill, C., Du, L., Johnson, M., and McCullough, B. D. (2024), Comparing programming languages for data analytics: Accuracy of estimation in Python and R, *WIREs Data Mining*

*Knowl Discov.* **2024**, e1531.

Klein, A., and Mélard, G. (2023b), An algorithm for the Fisher information matrix of a VARMAX process, *Algorithms* **16**, 364. https://doi.org/10.3390/a16080364. 14 pp.

McCullough B. D (1998), Assessing the reliability of statistical software: part I, *The American Statistician* **52**, 358-366.

McCullough B. D. (1999), Assessing the reliability of statistical software: part II, *The American Statistician* **53**, 149-159.

McCullough B. D. (2004), Some details of nonlinear estimation, in Micah Altman, Jeff Gill and Michael P. McDonald (ed.) *Numerical Issues in Statistical Computing for the Social Scientist*, Chapter 8. Wiley, New York, pp 199-218.

McCullough B. D., and Renfro C. R. (2000), Some numerical aspects of nonlinear estimation, *J Economic and Social Measurement* **26**, 63-77.

Mélard, G. (1984), Algorithm AS197: A fast algorithm for the exact likelihood of autoregressive-moving average models. *Journal of the Royal Statistical Society Series C, Applied Statistics* **33**, 104-114.

Mélard, G. (1985), Exact derivatives of the likelihood of ARMA processes, *1985 Proceedings of the Statistical Computing Section*, American Statistical Association, Washington D.C., pp. 187-192.

Mélard, G. (2024), Standard errors in time series and non-linear models, submitted.

Mélard, G. et Klein, A. (2023), Sur les algorithmes pour l'information de Fisher de modèles vectoriels dynamiques, 54es Journées de Statistique de la SFdS, JdS2023, Bruxelles, 3-6 juillet 2023.

Newbold, P., Agiakloglou, C., and Miller, J. (1994), Adventures with ARIMA software. *International Journal of Forecasting* **10**, 573-581.

Ogita, T., Rump, S. M., and Oishi, S. (2005), Accurate sum and dot product, *SIAM J. Scientific Computing* **26**, 1955-1988.

Yalta, A. T., and Jenal O. (2009), On the importance of verifying forecasting results, *International Journal of Forecasting* **25**, 62-73.

# Impact de la métrique pour le clustering de séries temporelles de quaternions: Application aux patients atteints de sclérose en plaques

Klervi Le Gall [1] & Lise Bellanger [2] & Aymeric Stamm[3] & David Laplaud [4]

[1] *Laboratoire de Mathématique Jean Leray, UMR CNRS 6629, Nantes Université, France.*
*klervi.legall@univ-nantes.fr*
[2] *Laboratoire de Mathématique Jean Leray, UMR CNRS 6629, Nantes Université, France.*
*lise.bellanger@univ-nantes.fr*
[3] *Laboratoire de Mathématique Jean Leray, UMR CNRS 6629, Nantes Université, France.*
*aymeric.stamm@cnrs.fr*
[4] *CR2TI, INSERM U1064, CHU de Nantes, Nantes Université, France.*
*david.laplaud@univ-nantes.fr*

**Résumé.** Le choix des métriques en classification non supervisée est une étape clé pour la construction de groupes séries temporelles de quaternions unitaires aux allures similaires. Nous détaillons ici la construction de plusieurs dissimilarités, ainsi la méthode de classification non supervisée par compromis choisie, et le cadre qui nous permet de d'évaluer ces partitions. Nous appliquerons cette méthodologie à un biomarqueur appelé Signature de Marche (SdM) qui caractérise la rotation de la hanche d'un individu au cours d'un cycle de marche moyen où les rotations sont représentées par des quaternions unitaires. Ces SdM ont été mesurées chez des patients atteints de sclérose en plaques pour lesquels nous obtenons une partition de données de marche interprétable et stable. La stabilité de cette méthode est étudiée à l'aide de l'inclusion de volontaires sains ainsi qu'avec la génération de données synthétiques.

**Mots-clés.** Séries temporelles de quaternions, donnés fonctionnelles, analyse de la marche, sclérose en plaques, classification non supervisée

**Abstract.** The choice of metrics in clustering is a key stage in the construction of clusters of unit quaternion time series with similar appearances. In this paper, we describe the construction of several dissimilarities, the clustering by compromise method chosen, and the framework that allows us to evaluate these clusters. We will apply this methodology to a biomarker called the Individual Gait Pattern (IGP), which characterises the rotation of an individual's hip during an average gait cycle, where rotations are represented by unit quaternions. These IGPs have been measured in multiple sclerosis patients for whom we obtain an interpretable and stable gait data partition. The stability of this method is studied by including healthy volunteers and by generating synthetic data.

**Keywords.** Quaternion time series, functional data, gait analysis, multiple sclerosis, clustering

# 1 Introduction

Les séries temporelles de quaternions unitaires sont des objets mathématiques complexes, pour lesquels il est nécessaire d'adapter les outils statistiques classiques. Nous savons que le choix de la métrique est un élément clé dans tous travaux de classification non supervisée, et que de nombreuses dissimilarités ou distances peuvent être construites entre deux séries temporelles de quaternions. L'objectif de ce travail est de proposer un cadre dans lequel nous pouvons appliquer une méthode de classification non supervisée à ces différentes métriques et ainsi évaluer leur pertinence pour créer des groupes interprétables et stables.

Le dispositif eGait permet de construire un biomarqueur appelé signature de marche (SdM). La SdM caractérise la rotation de la hanche d'un individu au cours d'un cycle de marche moyen. Les rotations sont représentées par des quaternions unitaires. La SdM fournit une mesure quantitative de la démarche à un moment donné sur laquelle nous travaillons afin de comprendre les troubles de la marche dans le contexte de la sclérose en plaques.

Dans le cadre du partitionnement de ces données nous souhaitons prendre en compte des informations externes, ou secondaires, qui viennent consolider les informations contenues dans les Signatures de Marche afin d'obtenir des groupes de patients aux démarches, et degrés de pathologie similaires. Ces informations sont apportées par le score EDSS et ce partitionnement est rendu possible par une méthode de classification non supervisée par compromis.

Nous détaillons ici la méthodologie nous permettant d'évaluer la pertinence des différentes métriques définies pour la classification non supervisée par compromis des séries temporelles de quaternions unitaires.

# 2 Matériel et méthodes

## 2.1 Données de marche

La démarche des patients est enregistrée à l'aide d'un système de capteurs qui transmet le vecteur d'orientation absolue du dispositif positionné à la hanche des patients sous la forme d'un quaternion unitaire toutes les 0,01 secondes, une série temporelle de quaternions unitaires est donc recueillie pour chaque patient. Ces données brutes sont ensuite segmentées en cycles de marche et la *signature de marche* (SdM) d'un individu correspond au centre des cycles de marche détectés, ce centre est obtenu avec une méthode de k-means alignement Sangalli et al. (2010) pour lequel le nombre de cluster est égal à 1. La série temporelle est ensuite exprimée en pourcentage de la durée totale. Ces SdM sont des séries temporelles de quaternions (QTS) unitaires et décrivent les rotations "moyennes" de la hanche pendant un cycle de marche. Les étapes de l'algorithme menant à la SdM d'un individus se trouvent dans l'article de Drouin et. al (2022a).

Les données sur lesquelles nous nous appuyons sont issues de la recherche eMSGait, une étude monocentrique non contrôlée sur la quantification des troubles de la marche chez les

patients atteints de sclérose en plaques. Cette étude vise à évaluer la possibilité de répartir les patients SEP inclus dans l'étude en groupes homogènes selon leurs caractéristiques. Entre août 2021 et août 2022, les patients âgés de 18 à 64 ans atteints de SEP, qui n'ont pas observé de poussée au cours des 5 semaines dernières semaines, capables de marcher avec canne (ou deux cannes) ou sans assistance, non opposés à l'étude et affiliés à un régime de sécurité sociale ont été inclus au CHU de Nantes.

L'étude comporte une visite unique correspondant à une consultation programmée dans le cadre de la prise en charge et du suivi habituel. Les données de marche ont été recueillies par le capteur de mouvement MetaMotionR r0.4 de MbientLab placé sur une ceinture au niveau de la hanche droite lors de deux tests neurologiques *Time 25 Foot Walk* (T25FW) Kieseier-Pozzilli (2012) espacés de 5 minutes. Des données cliniques décrivant la sévérité de leur pathologie ont également été mesurées par les cliniciens du Centre d'Investigation Clinique du CHU de Nantes.

Parmi les données cliniques mesurée figure le score EDSS (Expanded Disability Status Scale) décrit par Kurtzke (1983). Cet indicateur, évalué par un neurologue, permet d'attribuer un score entre 0 (examen neurologique normal) et 10 (Décès à cause de la SEP) aux patients. L'examen clinique est composé de sept ou huit fonctions neurologiques à évaluer qui peuvent impacter ou non la démarche des patients. Les patients peuvent alors être répartis en 3 groupes en fonction de la sévérité globale de leur pathologie représentée par leur score EDSS:

- Sévérité faible (SF) : EDSS $< 2$
- Sévérité modérée (SM) : $2 \leq$ EDSS $< 4$
- Sévérité élevée (SE) : $4 \leq$ EDSS

Ces groupes sont illustrés dans la figure 1.

## 2.2 Méthodes

### 2.2.1 Séries temporelles de quaternions unitaires

Un quaternion est un élément (w,x,y,z) $\in \mathbb{R}^4$, qui est une extension des nombres complexe et qui s'écrit comme suit:

$$\mathbf{q} = (w, x, y, z)^\top = w + ix + jy + kz \in \mathbb{R}^4, \tag{1}$$

où $i$, $j$ et $k$ respectent $i^2 = j^2 = k^2 = ijk = -1$, Voight (2005).

**Les quaternions unitaires** sont ceux pour lesquels $||\mathbf{q}|| = w^2 + x^2 + y^2 + z^2 = 1$ de sorte que leur norme soit égale à 1. Le groupe des quaternions unitaires $\mathbb{H}_u$ forme un groupe de Lie isomorphe au groupe unitaire spécial $SU(2)$ qui couvre deux fois le groupe des matrices de rotations en 3-Dimensions, Dijkhuizen (1994).

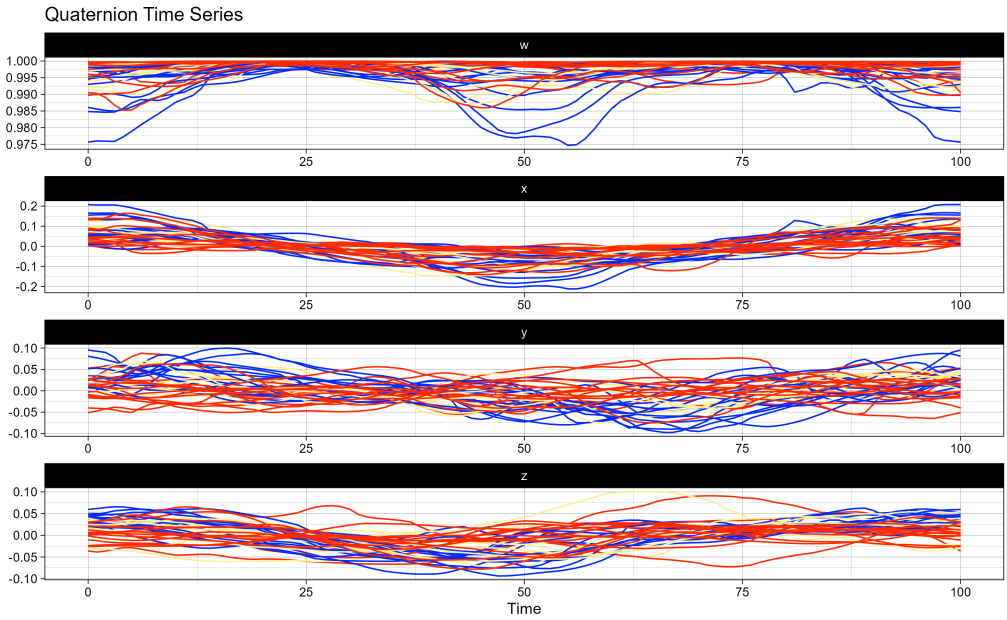



Figure 1: Signatures de Marche des 39 patients de l'étude eMSGait colorés par leur degré de sévérité pathologique, bleu: SF, jaune: SM, rouge: SE.

Les quaternions unitaires décrivent une rotation d'un angle $\theta$ autour d'un axe $\mathbf{u} = (u_x, u_y, u_z)^\top \in \mathbb{S}^2$ où $\mathbb{S}^2$ est la 2-sphère qui peut être exprimée comme suit:

$$\mathbf{q} = (cos\ \frac{\theta}{2}, \mathbf{u}_x\ sin\ \frac{\theta}{2}i, \mathbf{u}_y\ sin\ \frac{\theta}{2}j, \mathbf{u}_z\ sin\ \frac{\theta}{2}k)^\top \tag{2}$$

L'équation d'Euler pour les nombres complexes unitaires peut se généraliser généralise de la même manière aux quaternions unitaires. Il est alors possible d'écrire un quaternion unitaire sous sa *forme polaire*:

$$\mathbf{q} = \exp(\mathbf{u}\frac{\theta}{2}) = \cos\frac{\theta}{2} + \mathbf{u}\sin\frac{\theta}{2} \tag{3}$$

**Série temporelle de quaternion unitaires**

Une série temporelle de quaternions unitaires (QTS) est un ensemble de quaternions unitaires suivant une grille temporelle $t_{i,1}, \ldots, t_{i,n}$. On note une QTS comme: $\mathbf{Q}_i = (\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,n})$. Elle représente des rotations 3D consécutives dans le temps.

**Espace tangent d'une série temporelle de quaternions unitaires**

Nous pouvons déduire le *logarithme* d'un quaternion unitaire depuis sa forme polaire:

$$\ln(\mathbf{q}) = \ln(\exp(\mathbf{u}\frac{\theta}{2})) = \mathbf{u}\frac{\theta}{2} \tag{4}$$

Cette transformation est une application entre l'espace des quaternions unitaires $\mathbb{H}_u$ et l'espace tangent $\mathfrak{T}_{\mathbf{q}_0}(\mathbb{H}_u) \subseteq \mathbb{R}^3$ , au point $\mathbf{q}_0 = (1, 0, 0, 0)$, Piórek (2020). La transformation inverse est possible à l'aide de l'exponentielle:

$$\exp(\mathbf{q}) = \exp(w) \left( \cos(\|\mathbf{v}\|), \frac{\mathbf{v}^\top}{\|\mathbf{v}\|} \sin(\|\mathbf{v}\|) \right)^\top, \tag{5}$$

avec $\mathbf{v}$ la partie vectorielle de $\mathbf{q} = (x, y, z)^\top$.

### 2.2.2 Comment mesurer la distance entre Signatures de Marche?

La distance euclidienne dans $\mathbb{R}^4$ définit une métrique qui n'est pas fermée par rapport à l'algèbre des quaternions. Les quaternions unitaires vivent sur une sphère 3, donc la plus courte distance entre deux points correspond à une ligne géodésique, elle peut aussi être interprétée comme l'angle de rotation entre $\mathbf{q}_1$ et $\mathbf{q}_2$. La **distance géodésique** entre deux quaternions peut s'écrire:

$$d(\mathbf{q}_1, \mathbf{q}_2) = 2 arcos Re(\mathbf{q}_1^{-1} \mathbf{q}_2), \quad \mathbf{q}_1, \mathbf{q}_2 \in \mathbb{H}_u, \tag{6}$$

Elle a notamment été utilisée dans des travaux de Piórek-Jabłoński (2020).

Pour mesurer la distance entre deux signatures de marche, plusieurs options s'offrent à nous selon que l'on considère l'approche série temporelle ou l'approche fonctionnelle:

Il est possible d'utiliser une généralisation de l'algorithme de Dynamic Time Warping, fréquemment utilisée pour mesurer des dissimilarités entre séries temporelles, aux quaternions à l'aide de la distance géodésique comme propose Jabłoński (2012). Cette dissimilarité, s'écrit

$$QDTW(\mathbf{Q}_1, \mathbf{Q}_2) = \min_W \sum_{i=1}^{K} d(\mathbf{q}_{1,w_{i,1}}, \mathbf{q}_{2,w_{i,2}}), \tag{7}$$

Où $d(.,.)$ est la distance géodésique présentée dans l'équation 6, $W$ est une fonction d'alignement, et $K$ est le nombre d'alignements nécessaires entre $\mathbf{Q}_1$ et $\mathbf{Q}_2$.

Comme détaillé dans la section précédente, il est possible de projeter les séries temporelles de quaternions dans un espace tangent à l'hypersphère. Cette propriété nous permet d'utiliser des distances entre des fonctions, pour lesquelles on pourra choisir de faire de l'alignement fonctionnel ou non.

Soient $f_{\mathbf{Q}_1}$ et $f_{\mathbf{Q}_2}$ deux fonctions (deux séries temporelles de log-quaternions dans notre cas) correspondant aux projections de $\mathbf{Q}_1$ et $\mathbf{Q}_2$ dans l'espace tangent$\mathfrak{T}_{\mathbf{q}_0}(\mathbb{H}_u) \subseteq \mathbb{R}^3$, la **distance $L_2$ normalisée** s'écrit:

$$D_n(f_{\mathbf{Q}_1}, f_{\mathbf{Q}_2}) = \frac{\|f_{\mathbf{Q}_1} - f_{\mathbf{Q}_2}\|_2}{\|f_{\mathbf{Q}_1}\|_2 + \|f_{\mathbf{Q}_2}\|_2}, \tag{8}$$

Où $\|f_{\mathbf{Q}_1} - f_{\mathbf{Q}_2}\|_2 = (\int |f_{\mathbf{Q}_1} - f_{\mathbf{Q}_2}|^2)^{1/2}$

Une fonction peut $f_{\mathbf{Q}}$ également être caractérisée par sa *Square Root Slope Function* (SRSF) $g(t) \in L^2$, Tucker-Srivastava (2013)

$$g(t) = \frac{\dot{f}_{\mathbf{Q}}(t)}{\sqrt{\|\dot{f}_{\mathbf{Q}}(t)\|_2}}, \tag{9}$$

où $\dot{f}_{\mathbf{Q}}(t)$ est la dérivée en temps d'une log-série temporelle de quaternions.

La distance $D_s$ entre deux SRSF $g_1(t)$ et $g_2(t)$ est la norme $L_2$ de la différence. $D_s(f_{\mathbf{Q}_1}, f_{\mathbf{Q}_2}) = \|g_1(t) - g_2(t)\|_2$ .

Enfin, on peut décider d'aligner ces SRSF afin de calculer une distance entre les amplitudes des fonctions alignées. Soient $\gamma \in \Gamma(I)$ les fonctions nécessaires à l'alignement des SRSF, cette distance s'écrit:

$$D_y(f_{\mathbf{Q}_1}, f_{\mathbf{Q}_2}) = \underset{\gamma \in \Gamma}{inf} \|g_1 - (g_2 \circ \gamma)\sqrt{\dot{\gamma}}\|_2 \tag{10}$$

### 2.2.3 Classification non supervisée par compromis

#### Prétraitements des données

Les SdM sont calculées sur les patients dans des conditions précises et réplicables, en réalisant un test du Time-25-FootWalk (T25FW). Mais il est possible de considérer que ces données, comme un certain nombre des données issues de capteurs, peuvent être bruitées. Il faut alors se poser la question de débruiter ces données, par exemple, en appliquant une Analyse en Composante Principales Fonctionnelle Multivariée (MFPCA) sur les $f_{\mathbf{Q}}$ ,projections des SdM dans l'espace Tangent (TPCA), Ramsay-Silverman (2005) et Happ-Greven (2018). On choisit de reconstruire les données sur les $p$ premières composantes principales telles que l'on conserve environ 80 % de l'inertie du jeu de données.

De plus, certains patients atteints de SEP peuvent avoir une atteinte cognitive importante, qui ne se caractérise pas par une démarche particulière, mais ces patients peuvent avoir des difficultés à comprendre les cosignes du test de marche et à les appliquer. Pour ne pas biaiser les résultats, on peut exclure ces patients.

Avant de calculer les matrices de distances entre les observations du jeu de données,il y a donc quatre voies de prétraitements différentes à considérer:

- Suppression des patients fortement cérébraux,
- Suppression des patients fortements cérébraux et débruitage des SdM,

- Conservation des patients fortement cérébraux,
- Conservation des patients fortement cérébraux et débruitage des SdM.

**Une méthode de CAH par compromis: HclustCompro**

HclustCompro est une méthode de classification ascendante hiérarchique par compromis, qui prend en compte une source d'information principale et une source d'information secondaire et permet d'appliquer une CAH à une matrice de distance considérée comme le meilleur compromis entre ces deux sources d'informations, cette méthode a déjà été appliquée aux données de marche par Drouin et al (2022b).

La première étape de cette méthode est le choix des mesures de dissimilarités, chacunes des distances listée dans la section 2.2.2 est normalisée, puis sera choisie comme source d'information principale $D_1$, et la distance de Gower normalisée, qui mesure les différence entre les scores EDSS (ou score spécifique à la marche) des patients est choisie comme source d'information secondaire $D_0$.

On cherche alors le paramètre $\alpha$ qui détermine la matrice de distance qui serait la meilleure combinaison convexe de ces deux sources d'information $D_\alpha = \alpha \times D_1 + (1-\alpha) \times D_0$.

$D_\alpha$ est alors calculée pour chaque $\alpha$ sur une grille entre 0 et 1, et une CAH, avec stratégie d'agglomération *moyenne* ou *complète* est appliquée pour déterminer le dendrogramme $T_\alpha$.

La valeur de pondération telle que $T_\alpha$ résulte du meilleur compromis entre les deux sources de données initiales, $\hat{\alpha}$, est déterminée par la minimisation du critère suivant:

$$CorCrit_\alpha = |Cor(D_1, D_\alpha^{coph}) - Cor(D_0, D_\alpha^{coph})|, \tag{11}$$

où $D_\alpha^{coph}$ est la distance cophénétique entre deux observations du dendrogramme $T_\alpha$ obtenu par CAH avec la même stratégie d'agrégation.

Lorsque nous travaillons avec une métrique basée sur des données fonctionnelles, les courbes à l'intérieur des groupes doivent être alignées sur leur centroïde une fois le regroupement hiérarchique effectué. Donc pour chaque nombre de classe possible, les courbes sont alignées sur les centres de classes afin de déterminer une nouvelle matrice de distances (norme $L_2$ sans alignement) entre les observations sur laquelle on pourra calculer les indicateurs permettant le choix du nombre de classes.

**Choix du nombre de classes**

Après le calcul du dendrogramme par la méthode HClustCompro, il est nécessaire de déterminer le nombre de classes dans lesquelles seront réparties les observations. Les critères *Within Sum of Squares (WSS)* et Silhouette Rousseeuw (1987) sont calculés pour chaque nombre de classes $K \in \{2, \ldots, N-1\}$. Ainsi soit une partition $\{C_1, \ldots, C_K\}$, on calcule:

$$WSS(K) = \sum_{k=1}^{K} \sum_{i \in V_k} D^2(\mathbf{Q}_i, \tilde{\mathbf{Q}}_k), \tag{12}$$

ou $D^2$ correspond à la dissimilarité choisie entre les deux observations, $\tilde{\mathbf{Q}}_k$ au médoïde de la classe et $V_k := \ell \in [[1, n]] : Q_\ell \in C_k$.

Ainsi que

$$S_{Q_i}(K) = \frac{b_i - a_i}{max(a_i, b_i)}, \tag{13}$$

où $a_i = \frac{1}{N_k} \sum_{i,j \in V_k}^{N_k} D(Q_i, Q_j)$ est la distance moyenne entre une observation $Q_i$ et toutes les observations de sa classe et $b_i = \frac{1}{N_k} \sum_{i,j \in V_{k_2}}^{N_k} D(Q_i, Q_j)$ est la distance moyenne entre une observation $Q_i$ et toutes les observations de la classe la plus proche à laquelle elle n'a pas été affectée , on s'intéresse aux valeurs de: $S(K) = \frac{1}{N} \sum_{i=1}^{N} S_{Q_i}(K)$.

Pour les distances calculées sur les projections des SdM dans $R^3$, on peut facilement adapter ces indicateurs en utilisant les matrices de distances entre fonctions ou SRSF après réalignement sur centroïdes de classes.

$$WSS(K) = \sum_{k=1}^{K} \sum_{i \in V_k} D^2(f_{\mathbf{Q}_1}^{(aligned)}, \tilde{f_{\mathbf{Q}_k}}), \tag{14}$$

avec $V_k := \ell \in [[1, n]] : f_{\mathbf{Q}_\ell}^{(aligned)} \in C_k$ et $D$ la norme $L_2$ entre les fonctions alignées. Le critère silhouette s'adapte de manière analogue.

Le nombre de classes retenu est celui qui maximise le critère silhouette sans que cela ne soit au détriment du critère WSS.

## 2.3 Critères d'évaluation

### 2.3.1 Choix de la meilleure partition

Nous appliquons donc la méthode de CAH par compromis HclustCompro avec des stratégies d'agglomération moyennes ou complètes aux matrices de distances obtenues pour chacunes des quatre voies de prétraitement possibles ainsi que chaque distance présentée dans la section 2.2.2.

Une fois que le dendrogramme représentant le meilleur compromis entre ces deux sources d'information est déterminé, nous nous assurons que la *corrélation cophénétique* entre la matrice $D_\alpha$ et la matrice , $D_\alpha^{coph}$ est suffisante, une corrélation élevée signifie que le dendrogramme est une représentation fidèle de ces données. Les dendrogrammes pour lesquels cette corrélation n'est pas suffisante seront exclus.

Parmis les possibilités restantes, nous écartons également les partitions avec un WSS trop important, indicateur d'une grande variabilité intra-classes.

Par la suite, le *V de Cramer* est calculé entre la répartition des observation proposée et

les scores EDSS. Il est est définit ainsi:

$$V = \sqrt{\frac{\chi^2}{N \times \min(K-1, J-1)}} \in [0; 1], \tag{15}$$

Où, $N$ est le nombre d'observations, $J$ est le nombre de modalités de la variable étudié (EDSS), $K$ est le nombre de classes et $\chi^2$ est la statistique du test du Chi-deux d'indépendance. A l'issue de cette étape, nous ne conservons que des options pour laquelle il y a une bonne relation entre le score EDSS et la partition, ce qui est traduit par un V de Cramer proche de 1.

Enfin, la meilleure partition parmi celles restantes est choisie selon la taille des groupes, l'étendue du score EDSS dans les groupes, le nombre de singletons et l'avis du Professeur David Laplaud, chef du Service de Neurologie au CHU de Nantes.

### 2.3.2 Stabilité de la méthode de CAH par compromis

La stabilité de cette partition est vérifiée dans un premier temps en ajoutant des volontaires sains à l'effectif de patients SEP et étudiant l'impact sur les groupes.

La stabilité de la méthode de classification semi-supervisée est testée en utilisant des jeux de données synthétiques. Nous avons développé une méthode de génération de données synthétiques de marche basée sur une approche mêlant réduction de dimension et plus proches voisins qui nous permet de générer des Signatures de Marche et des EDSS synthétiques. En utilisant ces données, nous vérifions que les groupes créés partagent des caractéristiques similaires en termes d'indicateurs spatio-temporels et cliniques.

# 3 Résultats et perspectives

La méthodologie suivie a permis de mettre en lumière quel prétraitement de données et quelle métrique étaient les meilleurs choix pour nous amener à former des groupes de patients à la démarche similaire, et que nous pouvons expliquer cliniquement. Nous avons également montré la stabilité de cette partition.

A l'avenir, nous nous intéresserons également à la possibilité de considérer que la distance que nous utilisons soit invariante aux rotations. Le cadre mathématique nécessaire à cette démarche a été publié par Kurtek, S., Srivastava, A., Klassen, E., et al. (2012), et l'adaptation que nous envisageons nous permettra de considérer que deux séries temporelles de quaternions sont identiques lorsque que la seule différence entre les deux séries réside dans le choix du référentiel d'origine.

# Bibliographie


Dijkhuizen, M.S. (1994), The double covering of the quantum group SOq(3), *Proceedings of the Winter School "Geometry and Physics". Circolo Matematico di Palermo*, 37, pp. 47-57.

Drouin, P. et Stamm, A. et Chevreuil, L. et al (2022a), Gait impairment monitoring in multiple sclerosis using a wearable motion sensor, *Medical Case reports and Reviews*, 5, pp. 1-5.

Drouin, P. et Stamm, A. et Chevreuil, L. et al. (2022b), Semi-supervised clustering of quaternion time series : Application to gait analysis in multiple sclerosis using motion sensor data, *Statistics in Medicine*, 42(4), pp. 433–456.

Happ, C. et Greven, S. (2018), Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains, *Journal of the American Statistical Association*, 113:522, pp. 649-659.

Jabłoński, B. (2012), Quaternion dynamic time warping, *IEEE Transactions on Signal Processing*, 60, pp. 1174–1183.

Kieseier, B.C. et Pozzilli,C. (2012), Assessing walking disability in multiple sclerosis, *Multiple Sclerosis Journal* 18.7, pp. 914-924.

Kurtek, S., Srivastava, A., Klassen, E., et al.(2012), Statistical Modeling of Curves Using Shapes and Related Features, *Journal of the American Statistical Association*, 107(499), pp. 1152–1165.

Kurtzke, J.F. (1983), Rating neurologic impairment in multiple sclerosis : an expanded disability status scale (edss), *Neurology*, 33(11).

Piórek, M. (2019), Analysis of Chaos for Quaternion Time Series, *Analysis of Chaotic Behavior in Non-linear Dynamical Systems : Models and Algorithms for Quaternions*,Springer International Publishing, pp. 73-88.

Piórek, M. et Jabłoński, B. (2020), A Quaternion Clustering Framework, *International Journal of Applied Mathematics and Computer Science*, 30.1, pp. 133-147.

Ramsay, J.O. et Silverman, B.W. (2005), Principal components analysis for functional data, *Functional Data Analysis* ,Springer, pp. 147-172.

Rousseeuw, P.J. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.

Sangalli, L.M. et Secchi, P. et Vantini, S. et al (2010), k-mean alignment for curve clustering, *Computational Statistics & Data Analysis*, 54, pp. 1219-1233.

Tucker, J. D., Wu, W. et Srivastava, A. (2013), Generative models for functional data using phase and amplitude separation, *Computational Statistics and Data Analysis*, 61, pp. 50-66.

Voight, J. (2005), Quaternion Algebras, *Springer Nature*

# Modélisation et prévision des flux de patients dans les services d'urgence de la région Grand-Est

Laurie Sapia

*Laboratoire de Mathématiques de Reims (LMR), UMR CNRS 9008, Université de Reims Champagne-Ardenne, France, laurie.sapia@univ-reims.fr*

**Résumé.** Les tensions hospitalières dans les services accueil d'urgence (SAU) français se sont accentuées ces dernières années et sont largement médiatisées. Des facteurs d'ordres organisationnel, structurel ou conjoncturel peuvent expliquer cette situation et incitent à une compréhension et une anticipation des flux de passage dans les SAUs. La modélisation des flux de passage dans les services hospitaliers, y compris les SAUs, a fait l'objet de nombreuses publications ces dernières années, en France et à l'international. Dans le présent document, nous proposons un modèle additif généralisé des flux journaliers, incorporant des informations exogènes (tels que les périodes de vacances scolaires, la proximité d'événements populaires), dans lequel la tendance est une fonction polynomiale et la saisonnalité est une combinaison d'harmoniques. Ce modèle est ajusté individuellement pour chaque SAU du Grand Est à partir des séries des passages fournies par l'association Est-RESCUE. Les performances de prévision des flux sont évaluées puis comparées aux modèles SARIMA, constituant la référence communément utilisée dans la littérature.

**Mots-clés.** Flux de patients des services d'urgence, modèle SARIMA, modèle additif généralisé, séries temporelles, prédiction.

**Abstract.** Hospital tensions in the French emergency departments (ED) have increased and are widely publicized. Organizational, structural or cyclical factors can explain this situation and encourage an understanding and anticipation of the flows of patients in the EDs. The modeling of the flow of passage in hospital departments, including EDs, has been the subject of numerous publications in recent years, in France and internationally. In this paper, we propose a generalized additive model of daily flows incorporating exogenous information (such as school holiday periods, proximity to popular events) in which trend is a polynomial function and seasonality is a combination of harmonics. This model is adjusted individually for each ED of the French Grand Est area from the series of passages provided by the association Est-RESCUE. The flow prediction performances are evaluated and compared to the SARIMA models, which are the reference commonly used in the literature.

**Keywords.** Emergency departments' patients flow, SARIMA model, Generalized additive model, time-series, forecasting.

# 1 Introduction

Par définition, la tension hospitalière est une situation exceptionnelle qui se produit lorsqu'il y a un déséquilibre entre les moyens disponibles de la structure hospitalière et le flux de patients dans ses services. Les SAUs sont particulièrement concernés, comme illustré lors de la pandémie de Covid-19 puis ensuite par les restrictions d'accès à certains services pendant les périodes contraintes, comme lors des congés estivaux. La surpopulation dans les SAUs et ses conséquences sanitaires, sociales ou financières ont fait l'objet de nombreuses études, dans plusieurs pays ; voir par exemple [5], [3], [8], [14], [10], [15].

Ce phénomène peut être expliqué par divers facteurs. Certaines sont d'ordre conjoncturel, comme les vagues d'épidémie saisonnières qui engendrent un afflux de patients important sur de courtes périodes. D'autres sont d'ordre organisationnel, les SAUs ayant dû s'adapter à l'évolution réglementaire et ont fait face à des pénuries de ressources matérielles et humaines de plus en plus marquées. Enfin, des facteurs structurels, comme le vieillissement de la population, entraînant l'augmentation de l'afflux de personnes âgées, dont la prise en charge est notoirement plus complexe ; voir par exemple [11].

Dans ce contexte, la prévision des flux de patients dans les SAUs permet d'anticiper et d'optimiser les ressources à déployer. Une littérature abondante existe sur le sujet, dont on peut trouver une revue dans [6] et [7]. Les modèles ARMA/SARIMA ont été employés pour la modélisation des flux de patients dès 1988 [12], notamment pour les SAUs des hôpitaux de Lille [9], puis de Troyes [1]. Ils constituent encore aujourd'hui la référence auxquels les autres modèles sont comparés. L'utilisation de nombreux autres modèles a de fait été explorée, par exemple : [13] utilise des réseaux de neurones récurrents ; [2] utilise un lissage exponentiel automatisé pour une prédiction mensuelle des passages.

En France, un réseau d'observatoires régionaux des urgences (ORU) maille le territoire ; les ORUs ont pour mission « la collecte, l'analyse et le partage des données du périmètre des urgences et des soins de premier recours d'une région et disposant en son sein d'une expertise de médecine d'urgence ». Ils fournissent aux Agences Régionales de Santé puis aux établissements des rétrospectives chiffrées de l'activité des SAUs, entre autres missions. Pour cela, ils peuvent s'appuyer sur les résumés de passage aux urgences (RPU), synthèse administrative standardisée de la prise en charge de chaque patient se présentant dans un SAU, dont la remontée vers les ORUs est systématisée.

Dans ce travail, nous proposons un modèle additif généralisé (MAG) pour les flux de passage quotidiens dans les SAUs de la région Grand Est. Notre modèle incorpore des composantes de tendance et de saisonnalité, ainsi qu'une composante événementielle permettant de capturer l'influence des périodes de vacances ou d'événements ponctuels. L'approche addivite offre, en comparaison d'un classique modèle SARIMAX, une grande souplesse sur la forme de chaque composante. Plus précisément, nous considérons ici une tendance polynomiale et une saisonnalité construite comme la superposition des oscillations de plus grandes harmoniques associées aux données d'ajustement. Ce modèle est ajusté individuellement à chaque SAU sur la base de la série temporelle des passages quotidiens dans ce SAU sur une période de trois ans, de 2016 à 2018. Ces séries sont fournies par l'association Est-RESCUE, ORU attachée à la région Grand Est, dans le cadre d'un partenariat de recherche avec l'Université de Reims

Champagne-Ardenne, après agrégation à l'échelle journalière des RPUs.

La suite du document est structuré comme suit. La description formelle de notre MAG est présentée dans la partie 2. La procédure de sélection de modèle et d'ajustement univariée à chaque SAU est explicitée dans la partie 3. Enfin, les performances de prévision du MAG sont comparées à celles du modèle SARIMA dans la partie 4.

# 2 Modélisation des flux dans les SAUs

Introduisons les notations qui seront utilisées tout au long du document. L'ensemble des SAUs du Grand Est est noté $\mathcal{D}$. Nous désignons génériquement par $d$ tout élément de $\mathcal{D}$, c'est-à-dire tout SAU. Pour tout SAU $d$, la série temporelle des nombres quotidiens de patients pris en charge par le service au cours d'une période fixe de $\mathbb{T}$ jours, $\mathbb{T}$ étant un nombre entier positif, est notée $(x_{1:\mathbb{T}}^{(d)}) = (x_t^{(d)})_{t\in\{1,\dots,\mathbb{T}\}}$. Nous supposons implicitement que le jour $t = 1$ correspond au 1er janvier 2016, la première date avec des enregistrements disponibles dans notre ensemble de données. Les enregistrements disponibles s'arrêtent le 31 octobre 2018, ce qui correspond à $\mathbb{T} = \mathbb{T}_{\max} := 1015$ jours. Nous supposons que, pour tout $d \in \mathcal{D}$, la séquence $(x_{1:\mathbb{T}}^{(d)})$ correspond aux premières valeurs de la réalisation d'un processus stochastique $\mathbf{X^{(d)}} = (X_t^{(d)})_{t\in\mathbb{N}^*}$ défini sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, c'est-à-dire,

$$x_t^{(d)} = X_t^{(d)}(\omega), \quad t = 1, \dots, \mathbb{T},$$

où $\omega \in \Omega$. Ainsi, par modélisation des flux de patients, nous entendons proposer une description quantifiée du comportement stochastique des processus $\mathbf{X^{(d)}}$, $d \in \mathcal{D}$. Lorsqu'aucune attention particulière n'est accordée à un département donné, nous supprimons l'exposant $^{(d)}$ dans les notations précédentes. Par conséquent, $(x_t)_{t\in\{1,\dots,\mathbb{T}\}}$ désignera la série des nombres quotidiens de patients dans un service générique non spécifié. Nous considérerons souvent des sous-séries de $(x_t)_{t\in\{1,\dots,\mathbb{T}\}}$ ou $(X_t)_{t\in N^*}$. En particulier, pour tout $t_1, t_2 \in \{1, ..., \mathbb{T}_{max}\}$, $t_1 < t_2$, nous notons $x_{t_1:t_2} := (x_{t_1}, x_{t_1+1}, \dots, x_{t_2-1}, x_{t_2})$ la série des valeurs successives de $x_t$ entre $t_1$ et $t_2$.

Nous proposons ici un modèle additif généralisé – désigné par THSR dans la suite, en référence aux quatre composantes dont il est la somme – avec tendance polynomiale et saisonnalité harmonique. Il est défini comme la superposition de quatre processus. Précisément, nous supposons

$$X_t = T_t + H_t + S_t + R_t + \epsilon_t, \tag{1}$$

où

— $T = (T_t)_{t\in\mathbb{N}^*}$ modélise la tendance à moyen et long terme du processus par une fonction polynomiale de t. Précisément, étant donnés $P_0(t) = 1$, $P_1(t), \dots, P_K(t)$, $K + 1$ polynômes orthogonaux de degrés croissants $0, 1, \dots, K$, nous supposons que

$$T_t = T_t^{(K)}(\mathbf{a}) = \sum_{k=0}^{K} a_k\, P_k(t), \quad t \in \mathbb{N}^*,$$

où $K \in \mathbb{N}^*$, $\mathbf{a} := (a_0, a_1, \dots, a_K) \in \mathbb{R}^{K+1}$ sont des paramètres numériques ;

— $H = (H_t)_{t\in\mathbb{N}^*}$ modélise les effets locaux (dans le temps) des événements à des dates spécifiques, telles que les vacances et jours fériés, le jour de Noël, les célébrations nationales, etc. Nous avons divisé les dates spécifiques en deux parties : celles qui sont communes à tous les services d'urgence du Grand Est d'une part, et celles qui sont spécifiques à chaque service d'autre part. Précisément, pour chaque $d \in \mathcal{D}$,

$$H_t = H_t^{(d)}(\mathbf{b}) = \sum_{i\in\mathcal{C}} b_i \mathbb{1}_i(t) + \sum_{j\in\mathcal{S}^{(d)}} b_j \mathbb{1}_j(t), \quad t \in \mathbb{N}^*,$$

où $\mathcal{C}$ désigne l'ensemble des dates communes à tous les SAUs, par exemple les jours de vacances scolaires, jours fériés, etc. $\mathcal{S}^{(d)}$ désigne l'ensemble des dates spécifiques au SAU $d$, par exemple les événements sportifs ou culturels massifs, $\mathbb{1}_i(\cdot)$ est la fonction indicatrice de $i$, c'est-à-dire, $\mathbb{1}_i(t) = 1$ si $t \in \{i\}$ et 0 sinon, et $b_i, b_j$ sont des paramètres réels à estimer. Les coefficients $b_i$ peuvent être constants sur une certaine période, par exemple, constants sur une période de jours fériés ou des périodes de jours fériés similaires (comme les jours fériés d'automne, avec une période de deux semaines chaque année à la fin du mois d'octobre). Les dates regroupées dans une cellule ont la même valeur de paramètre $b_i$ ;

— $S = (S_t)_{t\in\mathbb{N}^*}$ modélise le comportement saisonnier du processus. Nous supposons que cette composante saisonnière est une superposition de fonctions harmoniques. Précisément,

$$S_t = S_t(\mathcal{F}_m, \mathbf{c}, \mathbf{d}) = \sum_{f\in\mathcal{F}_m} \left(c_f \cos(2\pi f\, t) + d_f \sin(2\pi f\, t)\right), \quad t = 1, \ldots, T,$$

où $\mathcal{F}_m$ est un sous-ensemble constitué de $m$ fréquences bien choisies parmi l'ensemble des fréquences $\left\{\frac{1}{T}, \frac{2}{T}, \ldots, \frac{\lfloor (T-1)/2 \rfloor}{T}\right\}$, et $c_f$ et $d_f$, $f \in \mathcal{F}_m$, sont des paramètres réels ;

— $R = (R_t)_{t\in\mathbb{N}^*}$ est un processus stochastique de type $ARMA(p, q)$ avec des hyperparamètres $p$ et $q$ choisis parmi les valeurs $\{0, 1, 2, 3, 4\}$ ;

— $(\epsilon_t)_{t\in\mathbb{N}^*}$ est un bruit blanc représentant le terme résiduel.

Rappelons que les modèles doivent être sélectionnés et ajustés, de manière séparée, à partir de la séquence des flux de chaque SAU.

En résumé, la famille des modèles THSR dépend des hyperparamètres $K \in \mathbb{N}^*$, le degré maximal de la fonction polynomiale qui modélise la tendance à moyen et long terme, $n_{\mathcal{C}}$ le nombre de périodes de jours fériés, $n_{\mathcal{S}}$ le nombre d'événements spécifiquement liés à chaque SAU, $\mathcal{F}_m$ le sous-ensemble d'harmoniques à conserver dans la partie saisonnière, et $(p, q)$ les hyperparamètres de la composante ARMA $R_t$. Les paramètres des modèles sont $\mathbf{a} := a_{0:K} \in \mathbb{R}^{K+1}$ les coefficients de la tendance polynomiale, $\mathbf{b} := b_{1:(n_{\mathcal{C}}+n_{\mathcal{S}})} \in \mathbb{R}^{n_{\mathcal{C}}+n_{\mathcal{S}}}$ les effets des jours fériés et des événements spécifiques, $(\mathbf{c}, \mathbf{d}) := (c_{1:m}, d_{1:m}) \in \mathbb{R}^{2m}$ les amplitudes des effets périodiques, et $\phi \in \mathbb{R}^{p+1}, \theta \in \mathbb{R}^{q+1}$ les paramètres du modèle $ARMA(p, q)$.

Dans la partie 4, les performances en prévision du modèle THSR seront comparées à celles d'un modèle SARIMA. La famille des modèles SARIMA est d'usage commun dans de nombreux domaines d'étude de séries temporelles. On ne rappelle pas ici la définition formelle de ces modèles (dont on pourra trouver un exposé rigoureux et complet, par exemple, dans [4]).

Retenons simplement que les modèles SARIMA dépendent d'un ensemble de sept hyperparamètres entiers p,d,q,P,D,Q,s, souvent notés $(p,d,q)(P,D,Q)_s$, où $s$ représente la période de saisonnalité du modèle, qui sera fixée à $s = 7$ dans le cas présent. Ici, nous restreignons $p,q \in \{0,1,2,3,4,5,6\}$, $d,D \in \{0,1\}$ et $P,Q \in \{0,1,2,3\}$. Les avantages et potentielles limites des THSR relativement aux SARIMA sont identifiées dans le tableau 1.

| THSR | SARIMA |
|---|---|
| • Prise en compte d'événements exogènes<br>• Flexibilité sur la forme des tendances à moyen et long terme permettant d'intégrer les connaissances d'experts.<br>• Modélisation simple de la saisonnalité multiple.<br>• Paramètres faciles à interpréter | • Effet saisonnier évolue dans le temps.<br>• C'est un modèle classique, qui peut être facilement reproduit dans d'autres langages de programmation. |

TABLE 1 – Caractéristiques formelles distinguant les modèles THSR et SARIMA.

# 3 Sélection de modèles, estimation des paramètres et prévisions

Le processus d'ajustement des modèles SARIMA et THSR aux flux de patients dans un SAU se déroule en deux étapes : premièrement, pour chaque valeur possible des hyperparamètres, les paramètres du modèle sont estimés à partir d'un ensemble d'entraînement de données. Ensuite, dans l'étape de sélection de modèle, le choix « optimal » des hyperparamètres est retenu. La définition de « optimal » dépend de la famille de modèles considérée. Deux types de critères sont pris en compte : a) la minimisation d'une fonction de coût comprenant un terme de pénalité croissant avec le nombre de paramètres du modèle (typiquement, l'opposé de la log-vraisemblance avec une pénalité AIC ou BIC), et b) la maximisation de la performance prévisionnelle estimée par une procédure de type validation croisée, basée sur des métriques de performance. Cette procédure en deux étapes est coûteuse en termes de calcul et nécessite de restreindre les valeurs possibles des hyperparamètres des modèles à un petit ensemble fini de valeurs. La méthode de sélection de modèle utilisée pour chaque famille est également identifiée ; ces méthodes de sélection de modèles sont explicitement décrites dans les sections 3.1 et 3.2. Pour évaluer la performance de prévision d'un modèle donné, à horizon de $h$ jours, nous utilisons le critère de l'erreur absolue moyenne pondérée WMAPE (Weighted Mean Absolute Percentage Error), exprimée en pourcentage, $\text{WMAPE} = \frac{\sum_{i=n+1}^{n+h} |\hat{y}_i - y_i|}{\sum_{i=n+1}^{n+h} |y_i|} * 100$, où $(y_i)_{i=n+1,n+h}$ sont les valeurs observées de la série temporelle de test et $(\hat{y}_i)_{i=n+1,n+h}$ sont les valeurs prédites par le modèle considéré, $n$ étant la taille de la série d'apprentissage.

## 3.1 Sélection de modèles SARIMA

Pour la sélection de modèle SARIMA, nous utilisons le critère classique BIC. Pour toute valeur des hyperparamètres $(p, d, q)(P, D, Q)_s$, le critère BIC du modèle SARIMA associé est calculé à partir de la séquence d'apprentissage $(x_{1:T_{train}})$, avec $T_{train} = 1008$. Le modèle ayant la valeur BIC minimale est alors retenu. Nous avons considéré pour les hyperparamètres des modèles SARIMA les valeurs $p, q \in \{0, 1, 2, 3, 4, 5, 6\}$, $d, D \in \{0, 1\}$, $P, Q \in \{0, 1, 2, 3\}$ et $s = 7$. Il est à noter que cette procédure de sélection de modèle, appliquée aux 59 services d'urgence du Grand Est, est chronophage. Les calculs ont été effectués via une procédure parallélisée en mémoire partagée sur un seul processeur avec 28 cœurs du centre régional de calcul intensif ROMEO [1], de l'Université de Reims Champagne-Ardenne, nécessitant un temps de calcul d'environ 15 heures.

## 3.2 Procédure de validation croisée séquentielle pour la sélection de modèles THSR

Pour estimer l'erreur de prévision théorique, à horizon de 7 jours, d'un modèle THSR (1) donné, nous proposons la méthode suivante de type validation croisée, bien adaptée au présent contexte de données temporelles, que nous appellerons validation croisée séquentielle (VCS). Contrairement aux méthodes de validation croisée classiques pour des données i.i.d, la méthode proposée prend en compte la dépendance temporelle et l'aspect séquentiel des données. Le modèle THSR ayant l'erreur de prévision estimée la plus faible sera alors choisi. Précisément, nous procédons de la manière suivante. La séquence $x_{1:\mathbb{T}}$, $\mathbb{T} = 1015$, est utilisée pour générer 25 couples $(z_{j,t}^{\text{train}}, z_{j,t}^{\text{test}})$, $j = 1, \ldots, 25$, de sous-séquences d'entraînement et de test. Plus précisément, nous fixons $T = 840$ et extrayons séquentiellement des sous-séquences de longueur $T$ de $x_{1:\mathbb{T}}$ en avançant de 7 jours, afin de définir

$$z_{j,1:T}^{\text{train}} = x_{1+7(j-1):T+7(j-1)}, \quad z_{j,T+1:T+7}^{\text{test}} = x_{T+7(j-1)+1:T+7j}, \quad j = 1, \ldots 25.$$

Nous avons choisi $T = 840$ pour deux raisons :

- cela permet de construire un nombre significatif $((\mathbb{T} - T)/7 = (1015 - 840)/7 = 25)$ de périodes de test de 7 jours pour évaluer la performance des modèles en matière de prévision à J+7 (en calculant la moyenne des mesures de performance sur les 25 périodes) ;
- $T = 840 = 2^3 \times 3 \times 5 \times 7$ possède de nombreux diviseurs, notamment 2, 3, 4, 5, 7, 8, de sorte que la transformée de Fourier rapide (FFT) soit rapide en terme de temps de calcul et couvre notamment les effets saisonniers avec des périodes à valeurs entières (en particulier la période hebdomadaire 7 et ses multiples).

Dans le paragraphe suivant, nous décrivons la méthode de sélection des hyperparamètres des modèles THSR. Elle est constituée de trois étapes imbriquées consistant à choisir successivement les hyperparamètres de chaque composante additive (tendance, saisonnalité, ARMA),

1. ROMEO est constitué de 117 serveurs équipés de processeurs Intel® Skylake 6132 cadencés à 2,6 GHz pour un total de 3276 coeurs, de 280 accélérateurs Nvidia P100 SXM2 interconnectées par la technologie NVlink et de mémoire DDR4 à 2667 MT/s pour capacité mémoire distribuée de 15,3 To.

en tenant compte des résultats de l'étape précédente et en ajustant au passage le choix des hyperparamètres pour les résidus de l'étape précédente :

1. Choix de $K$ : Pour toute période $j = 1, \ldots, 25$, en tenant compte des variables exogènes $\mathcal{C} \cup \mathcal{S}$, pour tout degré du polynôme $K = 0, 1, \ldots, K_{max}$, avec $K_{max} = 10$, nous estimons par la méthode des moindres carrées les paramètres du modèle additif $T_t^{(K)} + H_t$, à partir de la séquence d'apprentissage $(z_{j,t}^{\text{train}})_{t=1,\ldots,T}$, comme suit

$$(\widehat{\mathbf{a}}^{(j)}, \widehat{\mathbf{b}}^{(j)}) := \arg\min_{(\mathbf{a},\mathbf{b})} \sum_{t=1}^{T} \left( z_{j,t}^{\text{train}} - \sum_{k=0}^{K} a_k P_k(t) - \sum_{i \in \mathcal{C} \cup \mathcal{S}} b_i \delta_i(t) \right)^2 .$$

   Les valeurs prédites à J+7, selon ce dernier modèle, s'écrivent

$$\widehat{z}_{j,t}^{(K)} = T_t^{(K)}(\widehat{a}_{0:K}^{(j)}) + H_t(\widehat{\mathbf{b}}^{(j)}), \quad t = T+1, \ldots, T+7.$$

   L'erreur de prévision correspondante est donc estimée par

$$\mathcal{E}_1(K) = \frac{1}{25} \sum_{j=1}^{25} \frac{\sum_{t=T+1}^{T+7} |\widehat{z}_{j,t}^{(K)} - z_{j,t}^{\text{test}}|}{\sum_{t=T+1}^{T+7} |z_{j,t}^{\text{test}}|}.$$

   Le choix optimal du degré $K$ est défini alors comme suit

$$\widehat{K} := \arg\min_{K} \mathcal{E}_1(K).$$

2. Sélection des hyperparamètres de la composante $S_t$ : la sélection des hyperparamètres de cette composante, comme l'estimation de ces paramètres, se fera à partir des résidus d'ajustement, notés $y_t$, de la séquence $x_t$, par le modèle de tendance sélectionné à l'étape précédente, i.e.,

$$y_t = x_t - T_t^{(\widehat{K})}(\widehat{\mathbf{a}}^{(j)}) - H_t(\widehat{\mathbf{b}}^{(j)}), \quad t = 1, \ldots, T.$$

   Pour toute période $j = 1, \ldots, 25$, on calcule la transformée de Fourier discrète de la séquence

$$z_{j,1:T}^{\text{train}} = y_{1+7(j-1):T+7(j-1)}, \quad T = 840,$$

   i.e.,

$$h^{(j)}(k/T) = \frac{1}{T} \sum_{t=1}^{T} z_{j,t}^{\text{train}} \exp\left( -i 2\pi \frac{k}{T}(t-1) \right), \quad k = 0, \ldots, T-1.$$

   Nous calculons ensuite le spectre moyen, sur les 25 périodes, à chaque fréquence $k/T$, $k = 1, \ldots, \lfloor (T-1)/2 \rfloor = \lfloor 839/2 \rfloor = 419$, i.e., $h(k/T) = \frac{1}{25} \sum_{j=1}^{25} h^{(j)}(k/T)$, $k = 1, \ldots, 419$. L'ensemble des fréquences seront ordonnées suivant leur niveau d'énergie (dans l'ordre décroissant) $|h(k/T)|$, $k = 1, \ldots, 419$. Notons alors $f_1, \ldots, f_m$, les fréquences ordonnées, et $\mathcal{F}_m := \{f_1, \ldots, f_m\}$, $m = 1, \ldots, M$, la suite croissante d'ensembles de fréquences. Ceci permet de construire des modèles croissants pour la composante saisonnière $S_t$ :

$$S_t = S_t(m, \mathbf{c}, \mathbf{d}) = \sum_{i=1}^{m} \left( c_i \cos(2\pi f_i\, t) + d_i \sin(2\pi f_i\, t) \right), \quad t = 1, \ldots, T.$$

Le problème se ramène alors à la sélection du nombre optimal d'harmoniques $m$, $m = 1, \ldots, M = 40$. Pour tout $m = 1, \ldots, M$, et pour toute période $j = 1, \ldots, 25$, on estime les paramètres $\mathbf{c} := (c_1, \ldots, c_m)$ et $\mathbf{d} := (d_1, \ldots, d_m)$ par moindres carrés, i.e.,

$$(\widehat{\mathbf{c}}^{(j)}, \widehat{\mathbf{d}}^{(j)}) := \arg\min_{\mathbf{c},\mathbf{d}} \sum_{t=1}^{T} (z_{j,t}^{\text{train}} - \sum_{i=1}^{m} (c_i \cos(2\pi f_i t) + d_i \sin(2\pi f_i t)))^2.$$

Les valeurs prédites à J+7, selon ce modèle, sont données par

$$\widehat{z}_{j,t}^{(m)} = S_t(m, \widehat{\mathbf{c}}^{(j)}, \widehat{\mathbf{d}}^{(j)}), \quad t = T+1, \ldots, T+7.$$

L'erreur de prévision de ce modèle, à $m$ harmoniques, est estimée, en prenant la moyenne sur les 25 périodes, comme suit

$$\mathcal{E}_2(m) := \frac{1}{25} \sum_{j=1}^{25} \frac{\sum_{t=T+1}^{T+7} |\widehat{z}_{j,t}^{(m)} - z_{j,t}^{\text{test}}|}{\sum_{t=T+1}^{T+7} |z_{j,t}^{\text{test}}|}.$$

Enfin, le choix optimal du nombre d'harmonique $m$ est défini par

$$\widehat{m} := \arg\min_{m} \mathcal{E}_2(m).$$

3. Choix de modèle ARMA(p,q) : Nous utilisons le critère BIC pour calibrer un modèle $ARMA(p,q), p, q \in \{0,1,2,3,4\}$ à partir des séquences résiduelles

$$x_t - \left( T_t^{(\widehat{K})}(\widehat{\mathbf{a}}) + H_t(\widehat{\mathbf{b}}) + S_t(\widehat{m}, \widehat{\mathbf{c}}, \widehat{\mathbf{d}}) \right), \quad t = 1, \ldots, T.$$

Le modèle THSR retenu s'écrit alors

$$X_t = T_t^{(\widehat{K})}(\widehat{\mathbf{a}}) + H_t(\widehat{\mathbf{b}}) + S_t(\widehat{m}, \widehat{\mathbf{c}}, \widehat{\mathbf{d}}) + R_t(\widehat{\phi}, \widehat{\theta}), \quad t = 1, \ldots, T. \tag{2}$$

Le tableau 2 présente les hyperparamètres retenus pour les modèles THSR des sept SAUs de la Marne (la présentation des résultats a été limitée à ces derniers par souci de concision). En

| SAU | Chalons | Epernay | Vitry | Reims-ad. | Reims-ped. | Courlancy | Reims-B. |
|---|---|---|---|---|---|---|---|
| $\widehat{K}$ | 3 | 1 | 1 | 1 | 1 | 3 | 4 |
| $\widehat{m}$ | 4 | 2 | 11 | 10 | 11 | 6 | 29 |
| $(p,q)$ | (3,0) | (3,1) | (1,1) | (1,1) | (3,0) | (1,0) | (0,0) |

TABLE 2 – Les hyperparamètres retenus lors de la procédure de sélection du modèle THSR par validation croisée séquentielle pour les sept SAUs de la Marne ($\widehat{K}$ : degré du polynôme modélisant la tendance, $\widehat{m}$ : nombre d'harmoniques, $(p,q)$ : les hyperparamètres de l'ARMA)

regardant ce tableau, et de manière plus large tous les SAUs que nous avons à disposition, nous pouvons remarquer que la tendance est modélisée par un polynôme de degrés 1 ou 3, et dans plus de la moitié des SAUs, le polynôme de degrés 1 suffit. Concernant les harmoniques sélectionnées, on remarque des similitudes entre les SAUs. En effet, la saisonnalité 7 et ses multiples sont présents systématiquement, par exemple $7 * \frac{1}{2}$, $7 * \frac{1}{3}$, $7 * 30$, ... Les SAUs pour lesquels le nombre d'harmoniques retenues est important se distinguent des autres par la présence d'oscillations de basses fréquences. Enfin, les paramètres du modèles ARMA sont assez proches, $p = 1$ ou $p = 3$ et $q = 0$ ou $q = 1$, ce qui signifie qu'il n'y a plus beaucoup à modéliser ; dans certains cas, le terme résiduel $R_t$ est négligeable et peut être assimilé à un bruit blanc (e.g. SAU de Reims-Bezannes).

# 4 Comparaison des performances de prédictions

En reprenant les même SAUs que précédemment, nous résumons dans le tableau 3 les erreurs de prédictions estimées et les écarts-types des estimations, du modèle THSR ainsi que du modèle SARIMA. En général, le modèle SARIMA a une erreur de prédiction légèrement supérieure à celle du modèle THSR, ce qui s'illustre bien en faisant la moyenne sur les quelques SAUs sélectionnés. Nous obtenons 12.5% pour le modèle SARIMA et 12.1% pour le modèle THSR. Le terme "WMAPE moyenne" correspond à la moyenne sur les 25 fenêtres glissantes et le terme "WMAPE écart-type" désigne l'écart-type empirique des estimations correspondant aux 25 fenêtres glissantes. Concernant les écarts-types, nous remarquons qu'en moyenne sur tous les SAUs, la différence est très légère entre les 2 modèles. En effet, la moyenne sur tous les SAUs est de 3.58 pour les modèles THSR et de 3.63 pour le modèle SARIMA.

| | SAU | Chalons | Epernay | Vitry | Reims-ad. | Reims-ped. | Courlancy | Reims-B. |
|---|---|---|---|---|---|---|---|---|
| THSR | WMAPE moy | 10.1 | 11.3 | **13.3** | **6.8** | **11.0** | **15.3** | **16.6** |
| | WMAPE ecart-t. | 3.4 | 4.1 | **5** | **1.6** | **3.1** | **4.9** | 7.9 |
| SARIMA | WMAPE moy | **10** | **11.2** | 14.3 | 7.4 | 12.2 | 15.9 | 16.7 |
| | WMAPE écart-t. | **2.9** | **3.8** | 5.2 | 1.8 | 4.1 | 6.4 | **7.4** |

Table 3 – Comparaison des deux modèles THSR et SARIMA

Le tableau 4 présente les erreurs de prédictions estimées ainsi que les écarts-types empiriques des estimations associées au modèle THSR, en identifiant l'apport de chaque nouvelle composante du modèle, pour les mêmes SAUs que précédemment. Le modèle "Naïf" consiste à prédire systématiquement le nombre moyen de passages journaliers observés sur la période d'apprentissage, le terme "TH" désigne la modélisation uniquement par la tendance et les variables exogènes, "THS" ajoute au modèle précédent la composante saisonnière, enfin le terme "THSR" désigne le modèle complet proposé. Nous observons que l'erreur de prévision décroît à chaque fois que l'on ajoute une nouvelle composante au modèle précédent. Par exemple, en faisant la moyenne sur tous les SAUs disponibles, l'erreur de prévision passe de 11.8 à 11 puis à 10.9. Cependant, nous pouvons noter que l'ajout de la modélisation des résidus par l'ARMA n'améliore pas toujours le modèle. Nous observons également le même phénomène pour l'écart-type moyen qui passe de 3.89 à 3.69 puis à 3.58.

| | SAU | Chalons | Epernay | Vitry | Reims-ad. | Reims-ped. | Courlancy | Reims-B. |
|---|---|---|---|---|---|---|---|---|
| Naïve | WMAPE moy | 16.31 | 15.94 | 20.69 | 10.9 | 16.24 | 19.68 | 20.6 |
| | WMAPE ecart-t. | 8.55 | 7.91 | 8.53 | 3.22 | 7.03 | 7.19 | 8.47 |
| TH | WMAPE moy | 10.37 | 12.04 | 14.01 | 7.9 | 12.07 | 16.7 | 17.63 |
| | WMAPE ecart-t. | 3.67 | 4.54 | 4.27 | 1.91 | 4.45 | 6.05 | 8.7 |
| THS | WMAPE moy | 10.03 | 11.74 | 13.22 | 6.86 | 11.17 | 15.37 | 16.63 |
| | WMAPE ecart-t. | 3.44 | 4.57 | 5.1 | 1.63 | 3.3 | 4.82 | 7.91 |
| THSR | WMAPE moy | 10.09 | 11.31 | 13.3 | 6.79 | 11.01 | 15.28 | 16.63 |
| | WMAPE ecart-t. | 3.4 | 4.1 | 5.02 | 1.58 | 3.09 | 4.89 | 7.91 |

Table 4 – Comparaison des modèles imbriqués du modèle naïf au modèle THSR à 3 composantes

# Références


[1] Mohamed Afilal, Farouk Yalaoui, Frédéric Dugardin, Lionel Amodeo, David Laplanche, and Philippe Blua. Forecasting the Emergency Department Patients Flow. *Journal of Medical Systems*, 40(7) :175, June 2016.

[2] Jochen Bergs, Philippe Heerinckx, Benoît Depaire, and Sandra Verelst. Knowing what to expect, forecasting monthly emergency department visits : A time-series analysis. *International Emergency Nursing*, 2014.

[3] Adrian Boyle, Adrian Boyle, Kathleen Beniuk, Ian Higginson, Paul Atkinson, and Paul Atkinson. Emergency department crowding : Time for interventions and policy evaluations. *Emergency Medicine International*, 2012.

[4] Peter J Brockwell and Richard A Davis. *Time series : theory and methods*. Springer science & business media, 1991.

[5] Mathew Foley, Nizar Kifaieh, and William K. Mallon. Financial impact of emergency department crowding. *Western Journal of Emergency Medicine*, 2011.

[6] Jan G. De Gooijer, Panos M. Pardalos, and Rob J. Hyndman. 25 years of time series forecasting. *International Journal of Forecasting*, 2006.

[7] Muhammet Gul and Erkan Celik. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, 2018.

[8] Esther Hing and Farida Bhuiya. Wait time for treatment in hospital emergency departments : 2009, June 2019.

[9] Farid Kadri, Sondès Chaabane, Fouzi Harrou, and Christian Tahon. Modélisation et prévision des flux quotidiens des patients aux urgences hospitalières en utilisant l'analyse de séries chronologiques. In *7ème conférence de Gestion et Ingénierie des Systèmes Hospitaliers (GISEH)*, pages 1–8, Liège, Belgium, July 2014.

[10] Li Luo, Yong Luo, Yang You, Yuanjun Cheng, Yingkang Shi, and Renrong Gong. A mip model for rolling horizon surgery scheduling. *Journal of Medical Systems*, 2016.

[11] Bozena Mielczarek. Estimating future demand for hospital emergency services at the regional level. *null*, null.

[12] P. C. Milner. Forecasting the demand on accident and emergency departments in health districts in the trent region. *Statistics in Medicine*, 1988.

[13] Carlos Narciso Rocha and Fátima Rodrigues. Forecasting emergency department admissions. *Journal of Intelligent Information Systems*, 56(3) :509–528, June 2021.

[14] Benjamin C. Sun, Renee Y. Hsia, Robert E. Weiss, David S. Zingmond, Li-Jung Liang, Weijuan Han, Heather McCreath, and Steven M. Asch. Effect of emergency department crowding on outcomes of admitted patients. *Annals of Emergency Medicine*, 2013.

[15] Denny Yu, Renaldo C. Blocker, Mustafa Y. Sir, M. Susan Hallbeck, M. Susan Hallbeck, Thomas R. Hellmich, Tara N. Cohen, David M. Nestler, and Kalyan S. Pasupathy. Intelligent emergency department : Validation of sociometers to study workload. *Journal of Medical Systems*, 2016.

# Estimation non paramétrique de densité

# Asymmetric kernel density estimation of heavy tailed data with application to clustering

Yasmina ZIANE [1] & Nabil ZOUGAB [2] & Smail ADJABI [3]

[1,3] *Research Unit LaMOS, University of Bejaia, Operational Research Department, Faculty of Exact Sciences. Bejaia, Algeria.*
[2] *Research Unit LaMOS, University of Bejaia, Department of Electrical Engineering, Faculty of Technology. Bejaia, Algeria.*

**Résumé.** Dans ce travail, nous proposons d'estimer la fonction de densité des données à queue lourde avec un support non négatif. Comme les données à queue lourde se caractérisent par des observations rares dans la queue, nous proposons de subdiviser l'ensemble de données en deux sous-ensembles de densité forte et faible, en utilisant k-means, une méthode de classification non supervisée d'apprentissage machine. Pour cela, nous construisons un nouvel estimateur qui combine les deux sous-ensembles avec deux noyaux asymétriques BSPE et gamma. Cependant, le paramètre de lissage sera estimé par l'approche bayésienne adaptative, développée à l'aide des deux noyaux proposés et de la méthode classique UCV. Pour évaluer les performances de l'estimateur proposé, une étude comparative avec l'estimateur classique sur des données simulées et réelles est réalisée.

**Mots-clés.** Approche bayesienne, données à queue lourde, k-means, noyau BSPE, noyau gamma.

**Abstract.** In this work, we propose to estimate the density function of heavy tailed data with non-negative support. As the heavy tailed data are characterized by rare observations in the tail, we propose to classify them into two subsets with high and low density, using k-means method, an unsupervised machine learning classification method. To this end, we construct a new estimator that combines two asymmetric BSPE and gamma kernels. However, the smoothing parameter will be estimated by the adaptive Bayesian approach, developed using the two proposed kernels and the classical UCV method. A comparative study between the proposed estimator and the classical estimator on simulated and real data is performed to evaluate their performance.

**Keywords.** BSPE kernel, gamma kernel, bayesian approach, UCV, heavy tailed data, k-means.

## 1 Introduction

The estimation of the probability density of heavy-tailed data is very complex due to their specific characteristics, such as rare observations in the tail. Non-parametric probability

density estimation by the kernel method is one of the most important technique for understanding the properties of the data distribution. However, a good estimation of the density depends on the correct choice of its parameters, kernel $K$ and smoothing parameter $h$. The use of symmetric kernels in the case of asymmetric data estimation, causes a serious problem at the edges, which is the edge bias problem. Edge bias is due to the allocation of weights by the symmetric kernel outside the density support when smoothing is carried out near the boundary. For this reason, a family of asymmetric kernels has been proposed in the literature.

In this work, we are interested in estimating the density of heavy-tailed data with support $[0, \infty)$, as these are characterised by sparse observations in the tail, then we propose to classify the data into two clusters with high and low density, using unsupervised machine learning classification methods. A good clustering approach is one that provides high homogeneity within the cluster and heterogeneity between clusters (Xu and Wunsch, 2005), (Sharma and ShikhaRai, 2012). The type of cluster we are interested in is the one that aims at optimising a given merit function, which will lead to a good clustering, i.e. observations of the same similarity are grouped in the same cluster. Different algorithms are based on this principle, such as k-means, k-center clustering (Gonzalez, 1985), (Mohiuddin et al., 2020), (P and Miin-Shen, 2020), etc. The k-means method was used to divide the heavy tail data into two subsets with high (HDR) and low (LDR) density where we will associate the BSPE and Gamma kernels with the (HDR) and (LDR) regions respectively.

The smoothing parameter plays a key role in non-parametric kernel estimation, it controls the degree of smoothing. Several methods have been discussed in the literature. Classical methods which generates a global smoothing parameter, see for example (Saulo et al., 2013). However, a problem in using a global bandwidth is that the kernel methods often produce unsatisfactory results for complex or irregular densities. For this reason, several authors have been motivated to propose an estimator of the variable smoothing parameter, see for exemple, (Somé, 2022), (Ziane et al., 2015), (Yasmina et al., 2018) and (Belaid et al., 2016; Belaid et al., 2018) who have invested in the estimation of the variable smoothing parameter by the Bayesian approach in the asymmetric case and (Brewer, 2000); (Zhang et al., 2006) and (Zougab et al., 2014) for symmetric kernels.

The main objective of this work, is to improve the quality of the estimator, by proposing a new estimator that combines two classical estimators associated to each cluster (HDR and LDR) with the asymmetric BSPE and gamma kernels respectively. The smoothing parameter will be estimated adaptively by the Bayesian approach using the explicit forms determined by (Somé, 2022) with gamma kernel (LDR) and (Ziane et al., 2015) with BSPE kernel (HDR).

The rest of this paper is organized as follows. In Section 2, we give a brief review of S-PE and gamma kernels density estimator. In Section 3, we present the k-means classification method. In Section 4, we first introduce the proposed kernel density estimators based on BSPE and gamma kernels. Second, we show some properties of $\hat{f}_{BSPE-Gamma}$ (bias and variance). We adapt the adaptive bayesian for the choice of bandwidth in Section 5. The performance of the proposed estimator will be tested via real and simulated data sets in Section 6, while Section 7 concludes.

## 2 A brief review: BSPE and gamma kernels density estimator

In this section, we present a brief recall of the classical $\hat{f}$ estimator with BS-PE kernels and gamma kernel.
Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) continuous random variables with an unknown probability density function (pdf) $f$ on the support $[0, \infty)$. The kernel estimator based on asymmetric densities is expressed as

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i) \tag{1}$$

with kernels defined in table 1

Table 1: Kernels

| Distribution | Kernel |
|---|---|
| Gamma | $\frac{y^{\frac{x}{h}}}{\Gamma(1+\frac{x}{h})h^{1+\frac{x}{h}}} \exp\left(-\frac{y}{h}\right)$ |
| BS−PE | $\frac{\nu}{2^{\frac{1}{2\nu}}\Gamma(\frac{1}{2\nu})\sqrt{4h}} \left(\frac{1}{\sqrt{xy}} + \sqrt{\frac{x}{y^3}}\right) \exp\left(\frac{-1}{2h^{\nu}}\left(\frac{y}{x} + \frac{x}{y} - 2\right)^{\nu}\right)$ |

## 3 Data clustering with machine learning

Clustering is a common technique in statistical data analysis that is used in many fields. Data clustering is based on partitioning a dataset into clusters, where the elements of each cluster are similar to each other and different from other clusters. The case of unlabeled data, leads to an unsupervised machine learning classification. Several unsupervised classification algorithms have been proposed: hierarchical clustering algorithm, K-means algorithm, K-medoids algorithm, etc.
The objective of our work is to propose a subdivision of the data by unsupervised machine learning classification methods. As heavy-tailed data are characterised by low density at the tail, a partitioning of these data into two subsets with high and low density by the k-means method is interesting. The k-means clustering algorithm has the following steps:

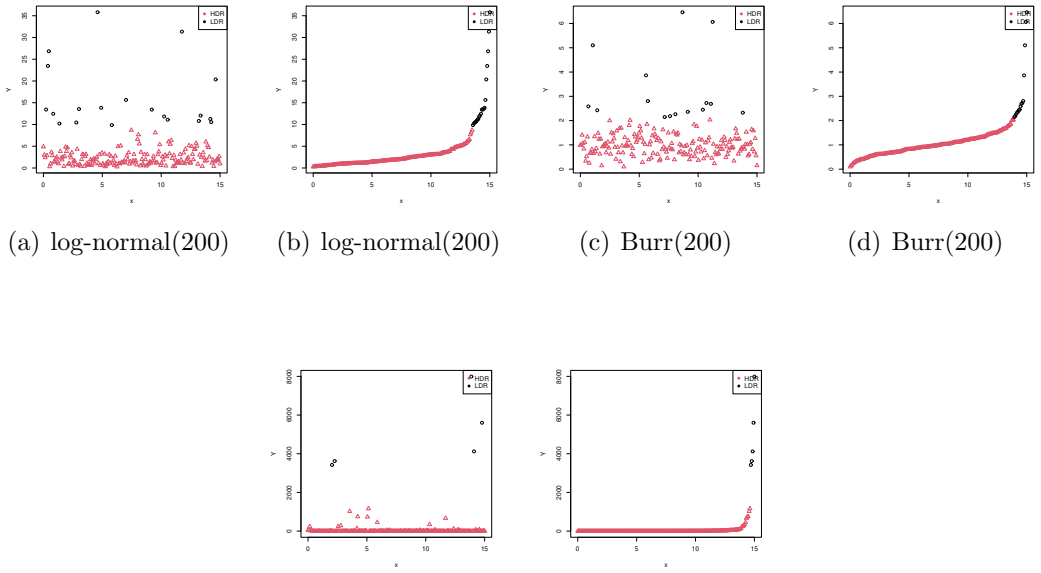
**a)** Define the number of clusters

**b)** Establish the centroid coordinates

**c)** Determine the distance between each observation and the centroid.

**d)** Grouping observations according to the minimum distance

In the case of heavy-tailed data, the number of clusters $k = 2$. Figure 1, illustrates the distribution of the data into two subsets of the different laws, which are characterised by a heavy tail (log-normal, burr and levy) and for sample sizes (n=200).




(a) log-normal(200) (b) log-normal(200) (c) Burr(200) (d) Burr(200)

(e) Levy(200) (f) Levy(200)

Figure 1: K-means repartition in high density region (HDR) and low density region (LDR).

From the graphs in figure 1, it can be seen that there is a low density at the extreme right (rare observations in the tail), in contrast to the extreme left where there is a high density.

# 4 The BSPE-Gamma kernel density estimation

We propose a new estimator of the density function, with a kernel function that would be flexible near zero and another kernel function that could estimate the tail of the density. Let $d$ be a threshold value that determines the proportion of the high density region in the sample space that is determined by the unsupervised machine learning classification method. The kernel density estimator is given by:

$$\hat{f}_{BSPE-Gamma}(x) = \begin{cases} \hat{f}_{(h,BSPE)}(x), & \text{if } x \in [0,d]; \\ \hat{f}_{(h,Gamma)}(x), & \text{if } x > d. \end{cases} \tag{2}$$

The bias and variance of the estimator $\hat{f}_{BSPE-Gamma}$ given by (2) are

$$biais\left(\hat{f}_{BSPE-Gamma}(x)\right) = \begin{cases} biais\left(\hat{f}_{(h,BSPE)}(x)\right), & \text{if } x \in [0,d]; \\ biais\left(\hat{f}_{(h,Gamma)}(x)\right), & \text{if } x > d. \end{cases} \tag{3}$$

$$biais\left(\hat{f}_{BSPE-Gamma}(x)\right) = \begin{cases} \frac{h\mu_1(g)}{2}\left(xf'(x) + x^2f''(x)\right) + o(h), & \text{if } x \in [0,d]; \\ h\left(f'(x) + \frac{1}{2}xf''(x)\right) + o(h), & \text{if } x > d. \end{cases} \tag{4}$$

$$Var\left(\hat{f}_{BSPE-Gamma}(x)\right) = \begin{cases} Var\left(\hat{f}_{(h,BSPE)}(x)\right), & \text{if } x \in [0,d]; \\ Var\left(\hat{f}_{(h,Gamma)}(x)\right), & \text{if } x > d. \end{cases} \tag{5}$$

$$Var\left(\hat{f}_{BSPE-Gamma}(x)\right) = \begin{cases} \frac{c^2}{C_{g^2} nh^{1/2}x} f(x) + o\left(\frac{1}{nh^{1/2}}\right), & \text{if } x \in [0,d]; \\ \begin{cases} \frac{1}{2\sqrt{\pi}} n^{-1}h^{-1/2}x^{-1/2} f(x) & \text{if } \frac{x}{h} \to \infty \\ \frac{\Gamma(2\kappa+1)}{2^{1+2\kappa}\Gamma^2(\kappa+1)} n^{-1}h^{-1} f(x) & \text{if } \frac{x}{h} \to \kappa \end{cases}, & \text{if } x > d. \end{cases} \tag{6}$$

where $\kappa$ is a nonnegative constant.

| $\mu_1(g)$ | $c$ | $C_{g^2}$ |
|---|---|---|
| $\frac{2^{1/nu}\Gamma(\frac{3}{2\nu})}{\Gamma(\frac{1}{2\nu})}$ | $\frac{\nu}{2^{1/2\nu}\Gamma(\frac{1}{2\nu})}$ | $2^{1/\nu}\Gamma(\frac{2\nu+1}{2\nu})\cos(\frac{\pi}{2\nu})$ |

# 5 Adaptive bayesian bandwidth selection

In this section, we present the adaptive smoothing parameters $(h_i)$ estimated by the adaptive Bayesian approach. The adaptive smoothing parameter represents the window $h$ estimated in each observation $X_i$ which gives a vector of the smoothing parameter $(h_1, h_2, \ldots, h_n)$. The explicit form of $h_i$ with the BSPE and gamma kernels developed respectively by (Ziane et al., 2015) and (Somé, 2022), are given by:

$$\hat{h}_{i,BSPE} = \frac{1}{\beta^{1/\nu}} \frac{\Gamma(\alpha - \frac{1}{2}\nu) \sum\limits_{j=1, i\neq j}^{n} \left(\frac{1}{\sqrt{x_i x_j}} + \sqrt{\frac{X_i}{X_j}}\right) \left[\frac{1}{2}\left(\frac{X_j}{X_i} + \frac{X_i}{X_j} - 2\right)^{\nu} + 1\right]^{-\alpha+\frac{1}{2\nu}}}{\Gamma(\alpha + \frac{1}{2}\nu) \sum\limits_{j=1, i\neq j}^{n} \left(\frac{1}{\sqrt{X_i X_j}} + \sqrt{\frac{X_i}{X_j}}\right) \left[\frac{1}{2}\left(\frac{X_j}{X_i} + \frac{X_i}{X_j} - 2\right)^{\nu} + 1\right]^{-\alpha-\frac{1}{2\nu}}} \tag{7}$$

$$\hat{h}_{i,Gamma} = \frac{1}{D_{ij}(\alpha,\beta)} \sum_{j=1, i\neq j}^{n} \left\{ \frac{(X_j + \beta) C_j(\alpha,\beta)}{\alpha} \mathbf{1}_{\{0\}}(X_i) + \frac{A_{ij}(\alpha,\beta) B_{ij}(\beta)}{\alpha - 1/2} \mathbf{1}_{(0,\infty)}(X_i) \right\} \tag{8}$$

where

$$\begin{aligned} A_{ij}(\alpha,\beta) &= \frac{\Gamma(\alpha+1/2)}{\beta^{\alpha} X_i^{1/2} \sqrt{2\pi} (B_{ij}(\beta))^{\alpha+\frac{1}{2}}}, \\ B_{ij}(\beta) &= X_i \log X_i - X_i \log X_j + X_j - X_i + \beta, \\ C_j(\alpha,\beta) &= \frac{\Gamma(\alpha+1)}{\beta^{-\alpha}(X_j+\beta)^{\alpha+1}}, \\ D_{ij}(\alpha,\beta) &= \sum_{j=1, i\neq j}^{n} \left(A_{ij}\mathbf{1}_{(0,\infty)}(X_i) + C_j \mathbf{1}_{(0)}(X_i)\right). \end{aligned}$$

$h_{i,BSPE}$ is the smoothing parameter vector associated to the high density region and $h_{i,Gamma}$ is the smoothing parameter vector associated to the low density region.

# 6 Simulation study

In this section, we present the simulation results obtained by evaluating the performance of the proposed estimator, compared to the classical estimator. This simulation study is based on samples of random variables simulated from heavy-tailed distributions, lognormal, burr and Levy presented in the table 2, for different sample sizes $n = 10, 50, 100$ and 500, and for a number of repetitions $N = 100$. The performances of the estimators are examined via integrated squared error (ISE) given by:

$$ISE := \int \left\{ \hat{f}(x) - f(x) \right\}^2 \tag{9}$$

where $\hat{f}$ is the BSPE-gamma, BSPE or gamma kernel estimators. the smoothing parameter $h$ is estimated by the adaptive Bayesian approach with the combination of BSPE-Gamma kernels which will be compared with the classical UCV approach.

Table 2: Distributions in the simulation study.

| | Distribution | Density | Parameters |
|---|---|---|---|
| **D1** | lognormal | $\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2} \left(\ln(x) - \mu\right)^2\right),\ x > 0$ | $(\mu, \sigma) = (1, 1)$ |
| **D2** | Burr | $\frac{kx^{k-1}}{(1+rx^k)^{r+1}},\ x > 0$ | $(k, r) = (3, 1)$ |
| **D3** | Levy | $\sqrt{\frac{c}{2\pi}} \frac{1}{(x-\mu)^{3/2}} \exp(-\frac{c}{2(x-\mu)}),\ x > \mu$ | $(\mu, c) = (0, 1/2)$ |

**remark:**
The performance of the adaptive Bayesian approach, depends on the choice of the a priori law parameters, used to develop the explicit form of the smoothing parameter (7) and (8). In this work, we followed the same principle as (Ziane et al., 2015) and (Somé, 2022).

The results of the comparative study between the proposed estimator and the classical BSPE and Gamma kernel estimator, with UCV for the bandwidth $h$ selection are presented in table 3, which reports averages of ISE and standard deviations values. From these we can observe:

- For the **D3** model, the proposed estimator is better than the classical estimator for all sizes of the sample.
- The proposed estimator outperforms the classical estimator for the **D1** and **D4** models for all sample sizes except for a large size $n = 500$.
- For the **D2** model, the classical estimator works best for medium and large sample sizes (100 and 500).

Table 3: Some expected values of ISE and their standard errors between parentheses based on 100 replications for the **D1**, **D2**, **D3** and **D4** distributions.

| $f$ | $n$ | $ISE_{BSPE-Gamma(UCV)}$ | $ISE_{BSPE(UCV)}$ | $ISE_{Gamma(UCV)}$ |
|---|---|---|---|---|
| | 10 | 0.0224 (0.0184) | 0.0331 (0.0229) | 0.0306 (0.0327) |
| **D1** | 50 | 0.0090 (0.0056) | 0.0094 (0.0062) | 0.0130 (0.0103) |
| | 100 | 0.0058 (0.0039) | 0.0061 (0.0053) | 0.0074 (0.0060) |
| | 500 | 0.0031 (0.0011) | 0.0017 (0.0012) | 0.0023 (0.0014) |
| | 10 | 0.0931 (0.0439) | 0.0984 (0.0807) | 0.1122 (0.0872) |
| **D2** | 50 | 0.0409 (0.0487) | 0.0260 (0.0169) | 0.0371 (0.0805) |
| | 100 | 0.0247 (0.0191) | 0.0160 (0.0116) | 0.0197 (0.0266) |
| | 500 | 0.0119 (0.0075) | 0.0068 (0.0046) | 0.0103 (0.0103) |
| | 10 | 0.0808 (0.0418) | 0.0918 (0.0441) | 0.1014 (0.0598) |
| **D3** | 50 | 0.0331 (0.0279) | 0.0341 (0.0358) | 0.0483 (0.0274) |
| | 100 | 0.0124 (0.0145) | 0.0159 (0.0141) | 0.0289 (0.0183) |
| | 500 | 0.0043 (0.0080) | 0.0053 (0.0042) | 0.0078 (0.0040) |
| | 10 | 0.0299 (0.0454) | 0.0485 (0.0456) | 0.0306 (0.0329) |
| **D4** | 50 | 0.0086 (0.0054) | 0.0147 (0.0163) | 0.0096 (0.0095) |
| | 100 | 0.0081 (0.0081) | 0.0090 (0.0069) | 0.0086 (0.0093) |
| | 500 | 0.0041 (0.0014) | 0.0029 (0.0021) | 0.0028 (0.0029) |

- If we compare the classical estimators with the BSPE and gamma kernels, we notice that the BSPE kernel, estimates the density better, in the case of models **D2**, **D3** and for model **D1** just for $n = 50$ and 100.

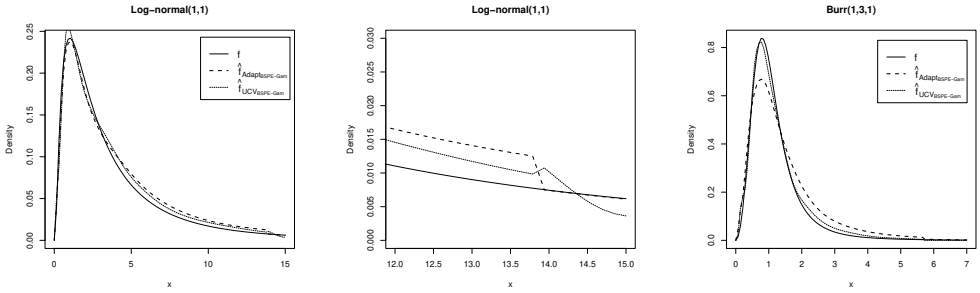
Table 4 presents averages ISE and standard deviations values resulting from comparison of the proposed estimator $\hat{f}_{BSPE-Gamma}$ with two smoothing parameter selection methods, the adaptive Bayesian approach and the classical UCV approach. From the results obtained, we can see that, the adaptive Bayesian approach is better than the classical UCV approach for all the models considered except **D2** for the sizes $n = 100$ and 500.

Table 4: Some expected values of ISE and their standard errors between parentheses based on 100 replications for the **D1**, **D2**, **D3** and **D4** distributions.

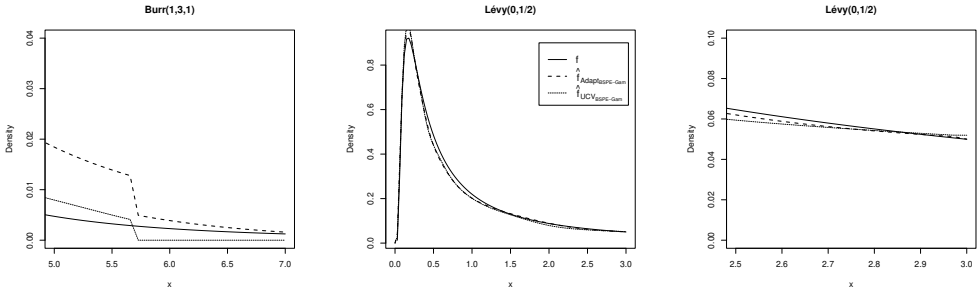
| $f$ | $n$ | $ISE_{BSPE-Gamma(bayes-adaptif)}$ (Std) | $ISE_{BSPE-Gamma(UCV)}$ (Std) |
|---|---|---|---|
| | 10 | 0.0215 (0.0176) | 0.0292 (0.0256) |
| **D1** | 50 | 0.0078 (0.0043) | 0.0122 (0.0057) |
| | 100 | 0.0047 (0.0028) | 0.0064 (0.0037) |
| | 500 | 0.0042 (0.0012) | 0.0056 (0.0023) |
| | 10 | 0.0873 (0.0355) | 0.1056 (0.0993) |
| **D2** | 50 | 0.0783 (0.0224) | 0.0934 (0.0360) |
| | 100 | 0.0580 (0.0184) | 0.0394 (0.0109) |
| | 500 | 0.0477 (0.0116) | 0.0348 (0.0057) |
| | 10 | 0.0774 (0.0458) | 0.1056 (0.0688) |
| **D3** | 50 | 0.0434 (0.0225) | 0.0452 (0.0463) |
| | 100 | 0.0100 (0.0134) | 0.0163 (0.0129) |
| | 500 | 0.0057 (0.0099) | 0.0057 (0.0038) |
| | 10 | 0.0246 (0.0206) | 0.0380 (0.0436) |
| **D4** | 50 | 0.0104 (0.0090) | 0.0193 (0.0331) |
| | 100 | 0.0064 (0.0031) | 0.0080 (0.0059) |
| | 500 | 0.0041 (0.0009) | 0.0057 (0.0024) |

In addition, We also compared the estimators graphically as shown in Figure 2, which plots the estimators of the lognormal, Burr, Levy and Weibull distributions for $n = 200$ and one replication, with Bayesian adaptive and UCV approaches for the choice of $h$. The left is the estimators of the distributions, and on the right is the zoom of the estimators tail. We note that the estimators are able to reproduce the uni-modality of the considered models. We can observe that the quality of smoothing is satisfactory, in particular at the tail of the

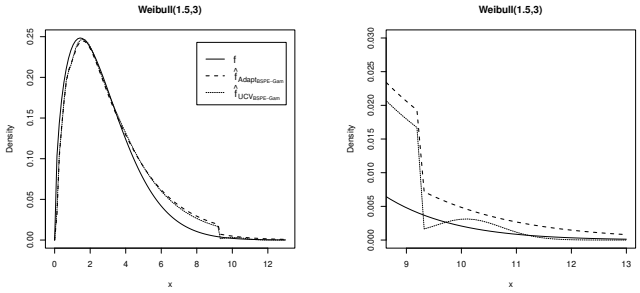
distributions, where the estimators estimate the tail well, whether it is with UCV or adaptive bayes approach.




(a) log-normal (b) log-normal (Tail) (c) Burr




(d) Burr (Tail) (e) Levy (f) Levy (Tail)




(g) Weibull (h) Weibull (Tail)

Figure 2: True pdf and kernel estimators for **D1**, **D2**, **D3** and **D4** with BSPE-gamma kernel, adaptive bayesian and UCV for $h$, for a sample size $n = 100$.

# 7 Real data

This section illustrates the performances of new estimators for real data set. These data present the vinyl chloride data obtained from clean upgrading, monitoring wells in mg/l; this

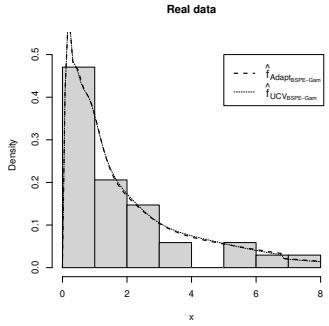



Figure 3: $\hat{f}_{BSPE-Gamma}$ estimator of vinyl chloride data with bayesian and UCV approaches

data set was used by (Ziane et al., 2021). Table 5 provides the descriptive summaries of vinyl chloride data.

Table 5: Descriptive summary of the vinyl chloride data set.

| Data set | $n$ | Max | Min | Median | Mean | SD | CS | CK |
|---|---|---|---|---|---|---|---|---|
| vinyl chloride | 34 | 8.000 | 0.100 | 1.150 | 1.879 | 1.952 | 1.603 | 5.005 |

Figure 3, shows the histogram and the $\hat{f}_{BSPE-Gamma}$ estimator with the adaptive Bayesian and UCV approaches, for the selection of the smoothing parameter on the real Vinyl chloride data. From this result, we can see that the estimators were able to reproduce the uni-modality and the tail of the data, the quality of the estimation is satisfactory with both considered approaches.

# 8 CONCLUSION

In this work, we proposed to estimate the density of data that are characterized by a heavy tail. Given this characteristic, we opted to divide the data set into two clusters (high density region (HDR) and low density region (LDR)) using the unsupervised k-means classification method. A new density estimator is proposed, associating to HRD the BSPE kernel and LDR the gamma kernel, the smoothing parameter is selected by the UCV and adaptive Bayesian approaches. A comparative study is performed to show the advantages of the proposed estimator over the classical estimator on simulated and real data.

# References

Belaid, N., Adjabi, S., Kokonendji, C. C., and Zougab, N. (2018). Bayesian adaptive bandwidth selector for multivariate discrete kernel estimator. *Communications in Statistics-Theory and Methods*, 47:2988–3001.

Belaid, N., Adjabi, S., Zougab, N., and Kokonendji, C. C. (2016). Bayesian bandwidth selection in discrete multivariate associated kernel estimators for probability mass functions. *Journal of the Korean Statistical Society*, 45:557–567.

Brewer, M. J. (2000). A bayesian model for local smoothing in kernel density estimation. *Statistics and Computing*, 10:299–309.

Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306.

Mohiuddin, A., Raihan, S., and Shamsul, I. S. M. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9:1295.

P, S. K. and Miin-Shen, Y. (2020). Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727.

Saulo, H., Leiva, V., Ziegelmann, F. A., and Marchant, C. (2013). A nonparametric method for estimating asymmetric densities based on skewed birnbaum–saunders distributions applied to environmental data. *Stochastic Environmental Research and Risk Assessment*, 27:1479–1491.

Sharma, S. and ShikhaRai (2012). Genetic k-menas algorithm implementation and analysis. *International Journal of Recent Technology and Engineering (IJRTE)*, 1(2):2277–3878.

Somé, S. M. (2022). Bayesian selector of adaptive bandwidth for gamma kernel density estimator on [0,): simulations and applications. *Communications in Statistics-Simulation and Computation*, 51:7287–7297.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE transactions on neural networks*, 16(3):645–678.

Yasmina, Z., Nabil, Z., and Smail, A. (2018). Birnbaum–saunders power-exponential kernel density estimation and bayes local bandwidth selection for nonnegative heavy tailed data. *Computational Statistics*, 33:299–318.

Zhang, X., King, M. L., and Hyndman, R. J. (2006). A bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, 50:3009–3031.

Ziane, Y., Adjabi, S., and Zougab, N. (2015). Adaptive bayesian bandwidth selection in asymmetric kernel density estimation for nonnegative heavy-tailed data. *Journal of Applied Statistics*, 42:1645–1658.

Ziane, Y., Zougab, N., and Adjabi, S. (2021). Body tail adaptive kernel density estimation for nonnegative heavy-tailed data. *Monte Carlo Methods and Applications*, 27(1):57–69.

Zougab, N., Adjabi, S., and Kokonendji, C. C. (2014). Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation. *Computational Statistics and Data Analysis*, 75:28–38.

# Kernel density estimation for a stochastic process with values in a Riemannian Manifold

Mohamed Abdillahi Isman [1,4] & Salah Khardani [2] & Papa Mbaye [1,3]
Wiem Nefzi [2] & Anne-Françoise Yao [1]

[1] *Laboratoire de Mathématiques Blaise Pascal, CNRS UMR 6620, Campus des Cézeaux, 63171 Aubière Cedex, France. anne.yao@uca.fr, mohamed.abdillahi_isman@doctorant.uca.fr*
[2] *Faculté des Sciences Tunis, Université El-Manar, Laboratoire de Modélisation mathématique, Statistique et Analyse Stochastique M2ϕSAS, Tunisie. khardani_salah@yahoo.fr, nefziwiem24@gmail.com*
[3] *Missioneo, Laboratoire de missioneo (Freeland Group), France. papa.mbaye@missioneo.fr*
[4] *Université de Djibouti, Djibouti. magouleh@gmail.com*

**Résumé.** Ce travail aborde la problématique de l'estimation de densité pour des observations ayant des valeurs dans une sous-variété riemannienne. Dans ce contexte, Pelletier (2005) a proposé un estimateur de densité à noyau pour des données indépendantes. Nous étudions le comportement asymptotique de l'estimateur de Pelletier lorsque les observations sont générées à partir d'un processus strictement stationnaire $\alpha$-mixing à valeurs dans cette sous-variété. En particulier, nous donnons les vitesses de convergence en termes d'erreur quadratique moyenne, en probabilité et presque sûrement. Nous établissons également un théorème central limite et illustrons notre propos à travers des simulations ainsi qu'une application à des données réelles.

**Mots-clés.** Estimateur de densité à noyau, Variétés riemanniennes, Condition de mélange, Processus stochastique, Consistance faible et forte, Théorème central limite.

**Abstract.** This paper is related to the issue of the density estimation of observations with values in a Riemannian submanifold. In this setting, Pelletier (2005) has proposed a kernel density estimator for independent data. We investigate here the behavior of Pelletier's estimator when the observations are generated from a strictly stationary $\alpha$-mixing process with values in this submanifold. In particular, we study the pointwise as well as the uniform weak and strong consistency of the estimator. Namely, we give the rate of convergence in mean square error meaning, in probability and almost surely. We also give a central-limit theorem and illustrate our purpose through some simulations and a real data application.

**Keywords.** : Kernel density estimator, Riemannian manifolds, Mixing condition, Stochastic process, Weak and Strong consistency, Central Limit Theorem.

## Motivation

The motivation behind this work is that the probability density estimator is crucial in many fields, including statistics, signal processing, machine learning, etc. The problem of estimating an unknown density using a kernel approach has been widely addressed when the

variable of interest lies in a Euclidean space, both for independent data and dependent data. However, in many applications , such as biology, spatial statistics, geology, image analysis, and medicine, the Euclidean assumption about the underlying geometry of the observations fails. An alternative is then to treat the data as lying on some submanifold.

# 1 Framework and the kernel of interest

Let $(X_t, t \in Z)$ defined on a probability space $(\Omega, \mathcal{A}, P)$ with values on Riemannian submanifold, $(\mathcal{M}, g)$ of $R^d$ $(d \geq 2)$ such that $X_t$'s are dependent and distributed as a random variable, $X$ with an unknown density $f$ on $\mathcal{M}$. We assume that $(X_t)$ satisfies an $\alpha-$mixing condition such that for any integer $n \geq 1$,

$$\alpha(n) = \sup_k \sup_{A \in \mathcal{F}_1^k(X),\, B \in \mathcal{F}_{k+n}^{\infty}(X)} \{|P(A \cap B) - P(A)P(B)|\}$$

where $\mathcal{F}_i^k(X)$ is the $\sigma-$field generated by $\{X_i,\ i \leq j \leq k\}$ and $\alpha(n)$ tends to zero.

In this work, we aim to study the behavior of the kernel estimator of Pelletier:

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n^d} \frac{1}{\theta_x(X_i)} K\left(\frac{d(x, X_i)}{h_n}\right),$$

based on observations of the process $(X_t, t \in Z)$, $X_i$, $i = 1, ..., n.$ . Here, $h_n$ is the bandwidth, $K$ is the kernel function, and $n$ is the number of observations. Then, we give some asymptotic results based on the following assumptions. Note that $C$ will represent a positive constant with no impact on the results.

**Assumptions**

**H1:** $K : R^d \to R_+$ is a bounded and continuous map such that

1. $K$ satisfies a Lipschitz condition.
2. $supp\ K = [0; 1]$.
3. $\int K(\|x\|)dx = 1$.
4. $\int xK(\|x\|)dx = 0$ or at least the vector, $\int xK\left(\|x\|\right)dx$ is orthogonal to span$\{\text{grad}f(x)\}$.
5. $\int \|x\|^2 K(\|x\|)dx < \infty$.

**H2:** The mixing coefficient of $X_i$ satisfies $\alpha(n) \leq Cn^{-\nu}$ for some $\nu > 2$.
**H3:** The bandwidth is such that $h_n \to 0$, $nh_n^d \to \infty$ and $h_n < \frac{h^*}{2}$ as $n \to \infty$.
**H4:** $f$ is bounded, twice continuously differentiable at any $x \in \mathcal{M}$ and $\|\text{Hess}f\,(x)\,\|_{HS} < \infty$, where $\|\cdot\|_{HS}$ is the Hilbert Schmidt norm (one can also consider the uniform norm of Hess$f\,(x)$).

**H5:** $\forall i, j$, the joint density $f_{i,j}$ of $(X_i, X_j)$ exists is such that

$$\sup_j \sup_{u,v \in \mathcal{M}\times\mathcal{M}} |f_{i,j}(u,v) - f(u)f(v)| < M,$$

for some $M > 0$.

### Some remarks on these assumptions

Since this work is related to situations where the data are both with values in a Riemannian manifold and dependent, the above assumptions appear as a mixed of conditions in both cases to derive consistency results for kernel density estimators in this setting. Namely,

1. The hypothesis **H2** and **H5** are classical when dealing with the study of the kernel density estimator for dependent data.
2. Assumption **H3** is the Riemannian manifolds counterpart of the classical assumptions on the bandwidth for dependent or independent data.
3. Assumptions **H1** and **H4** are classical regularity conditions on $K$ and $f$, helpful to control the bias terms (1) obtained using the Taylor expansion:

$$f(\exp_x(u)) = f(x) + < \text{grad}f(x), u > + \frac{1}{2}\text{Hess}f(x)(u,u) + o(\|u\|^2)$$

for all $u \in T_x\mathcal{M}$. Thus, for all $h > 0$ and $\|v\| \leq 1$, we have

$$f(\exp_x(hv)) = f(x) + h < \text{grad}f(x), v > + \frac{1}{2}h^2\text{Hess}f(x)(v,v) + o(h^2),$$

where grad and Hess denote the gradient and the Hessian operators respectively.

## 2 Some Theoretical Results

The assumptions stated above allow us to give some theoretical results on the asymptotic behavior of $\widehat{f}_n(x)$ that allow for evaluating the quality of the estimator in terms of bias and variance. Specifically, we study both the weak and strong consistency of $\widehat{f}_n(x)$ in terms of Mean Squared Error(MSE), probability and almost sure convergence.

### Weak and Strong Consistency

#### The bias term

The asymptotic behavior of the bias term has been studied in Pelletier (2005). However, we give here an explicit expression

$$b(x) := E\hat{f}_n(x) - f(x) = \frac{h_n^2}{2}\int_{B(1)} K(\|v\|)\text{Hess}f\,(x)\,(v,v)dv + o(h_n^2) \qquad (1)$$

Unlike the bias, all the results below are specific theoretical contributions of this paper.

**Asymptotic behavior of the variance of the estimator**

Under the assumptions **H1, H2** and **H5**, the variance of $\widehat{f}_n(x)$ is given by

$$V\left(\widehat{f}_n(x)\right) := E\left[\left(\widehat{f}_n(x) - E\widehat{f}_n(x)\right)^2\right] = \frac{1}{nh_n^d}\left[f(x)\int_{B(1)} K^2(\|v\|)dv + o(1)\right]$$

and then

$$nh_n^d V\left(\widehat{f}_n(x)\right) \underset{n\to\infty}{\longrightarrow} f(x)\int_{B(1)} K^2(\|v\|)dv.$$

**Rate of convergence in Mean Squared Error**

For each $x \in \mathcal{M}$, we set

$$MSE(x) := E\left((\widehat{f}_n(x) - f(x))^2\right) = b^2(x) + V\left(\widehat{f}_n(x)\right).$$

Under the assumptions **H1** to **H5**,

$$MSE(x) \leq C_{x,f} \times \left(h_n^4 + \frac{1}{nh_n^d}\right),$$

and since $\mathcal{M}$ is compact, and $f$ and $K$ are bounded (see **H1** and **H4**), we have

$$\sup_{x\in\mathcal{M}} MSE(x) \leq C \times \left(h_n^4 + \frac{1}{nh_n^d}\right).$$

An optimal rate in $MSE$ meaning for $\widehat{f}_n(x)$ is deduced from the previous result in the following.

The bandwidth which minimizes the function $x \mapsto MSE(x)$, under the assumption **H1** is given by:

$$h_{n,opt} = C_0\, n^{\frac{-1}{d+4}}$$

and the corresponding $MSE$ is

$$MSE(x) = C_{x,f}\, n^{\frac{-4}{4+d}} + o\left(n^{\frac{-4}{d+4}}\right)$$

where $C_{x,f} = C_0^4 C_1 + C_0^{-d} C_2$ with $C_0 = \left(\frac{dC_2}{4C_1}\right)^{\frac{1}{4+d}}$, $C_1 = \left(\int_{B(1)} \mathrm{Hess} f(x)(v,v)K(\|v\|)dv\right)^2$ and $C_2 = f(x)\int_{B(1)} K^2(\|v\|)dv$.

**Rate of Convergence: In Probability and Almost Surely**

We now give some pointwise as well as uniform rates of convergence in probability with some additional conditions.

Under the assumptions **H1** to **H5**, if $\alpha(n) \leq Cn^{-\nu}$ with $\nu > 2$ (as stated in **H2**) and if $\frac{nh_n^{d\frac{\nu+1}{\nu-1}}}{\log n} \to \infty$, then for a given $x \in \mathcal{M}$

$$|\widehat{f}_n(x) - f(x)| = O_p\left(h_n^2 + \sqrt{\frac{\log n}{nh_n^d}}\right).$$

Moreover, since $\mathcal{M}$ is compact, if $\nu > d+1$ and $n^{-1}(\log n)h_n^{-d\frac{\nu+d+3}{\nu-d-1}} \to 0$, we have

$$\sup_{x\in\mathcal{M}}|\widehat{f}_n(x) - f(x)| = O_p\left(h_n^2 + \sqrt{\frac{\log n}{nh_n^d}}\right).$$

As previously, with some additional conditions, we get the following almost surely rates of convergence.

Under the assumptions **H1** to **H5**, if $\alpha(n) \leq Cn^{-\nu}$ with $\nu > 3$ and $nh_n^{\frac{d(\nu+1)}{\nu-3}} (\log n)^{-\frac{\nu-1}{\nu-3}} g(n)^{\frac{-2}{\nu-3}} \to \infty$ with $h_n < \frac{h^*}{2}$ and $g(n) = \log n (\log\log n)^{1+\epsilon}$ for some small $\epsilon > 0$ then for each given $x \in \mathcal{M}$, we have

$$\widehat{f}_n(x) - f(x) = O_{a.s.}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n^d}}\right)$$

and because $\mathcal{M}$ is compact, we have

$$\sup_{x\in\mathcal{M}}|\widehat{f}_n(x) - f(x)| = O_{a.s.}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n^d}}\right).$$

### Asymptotic normality

We now state a Central Limit Theorem (CLT) for each given $x \in \mathcal{M}$.

Under the assumptions **H1**, **H2**, **H3** and **H5** , we have:

$$\sqrt{nh_n^d}\left(\widehat{f}_n(x) - f(x)\right) \longrightarrow \mathcal{N}\left(0, \sigma^2(x)\right),$$

where $\sigma^2(x) = f(x)\int_{B(1)} K^2\left(\|x\|\right) dx$.

## 3 Numerical results

During the talk, we will illustrate these theoretical results through some simulations and a real data application.

# 4 Some References

le Brigant, A. and Puechmorel, S. (2019), Approximation of densities on Riemannian manifolds, *Entropy,* 21(1).

Pelletier, B. (2005), Kernel density estimation on Riemannian manifolds, *Statistics & Probability letters*, 73(3):297 - 304.

García-Portugués et al. (2020), On optimal tests for rotational symmetry against new classes of hyperspherical distributions, *Journal of the American Statistical Association*, 115(532):1873 - 1887.

Cleanthous, et al. (2020), Kernel and wavelet density estimators on manifolds and more general metric spaces, *Bernoulli,* 26(3):1832 - 1862.

Berenfeld, C. and Hoffmann, M. (2021), Density estimation on an unknown submanifold, *Electronic Journal of Statistics,* 15(1):2179 - 2223.

Aamari, E. and Levrard, C. (2019), Non-asymptotic rates for manifold, tangent space and curvature estimation, *The Annals of Statistics,* 47(1):177 - 204

.

# KERNEL DENSITY ESTIMATION FOR CONTINUOUS TIME PROCESSES ON RIEMANNIAN MANIFOLD

Djack Guy-Aude Kouadio [1,2] , Vincent Monsan [1] & Anne-Françoise Yao [3,4]

[1] *Laboratoire de Mathématiques Informatique et Mécanique, Université Félix Houphouet Boigny, Abidjan - Côte d'Ivoire,*
*monsanv@gmail.com and guyaude.kouadio@ulagunes.com*
[2] *Université des Lagunes, Abidjan - Côte d'Ivoire,*
*guyaude.kouadio@ulagunes.com*
[3] *Laboratoire de Mathématiques Blaise Pascal, CNRS UMR 6620, Campus des Cézeaux, 63171 Aubière Cedex, France,*
*anne.yao@uca.fr*
[4] *Centre de Mathématiques Appliquées, Ecole polytechnique Route de Saclay 91128 Palaiseau Cedex,*
*anne.yao@uca.fr*

**Résumé.** Dans ce travail, nous nous intéresserons à l'estimation de la densité marginale, d'un processus stationnaire à temps continu $(X_t, t \in \mathbb{R})$, où les $X_t$ sont de même loi qu'une variable aléatoire $X$ à valeurs dans une sous variété Riemannienne, $(\mathcal{M}, g)$, de $\mathbb{R}^d$ $(2 \leq d < \infty)$ muni d'une mesure $\mu_g$ et $(\mathcal{M}, d_g)$ est complète, où $d_g$ est la métrique induite par $g$. Désignons par $f$ la densité de $X$ par rapport à $\mu_g$, que nous estimerons par une methode non paramétrique. L'estimateur à noyau proposé généralise d'une part l'estimateur classique pour les processus à temps continu à valeurs dans un espace euclidien aux sous-variétés Riemanniennes (voir par exemple Bosq, D. (1998)), et d'autre part l'estimateur proposé par Pelletier, B. (2005) pour des observations indépendantes et identiquement distribuées (*i.i.d.*) aux processus à temps continu. Lors de notre exposé, apres avoir présenté l'estimateur nous donnerons des résultats de convergence sur le comportement asymptotique de celui-ci obtenus sous conditions de mélange.

**Mots-clés.** *Estimateur à noyau d'une densité, variétés riemanniennes, conditions de mélange, processus stochastique à temps continu.*

**Abstract.** This work deals with the estimation of the marginal density of stationary continuous time process $(X_t, t \in \mathbb{R})$, where the $X_t$'s are distributed as a random variable $X$ with values in a Riemannian submanifold, $(\mathcal{M}, g)$, of $\mathbb{R}^d (2 \leq d < \infty)$ endowed with a measure $\mu_g$ and $(\mathcal{M}, d_g)$ is complete, where $d_g$ is the metric induced by $g$. Let $f$ be the density of $X$ with respect to $\mu_g$ which we will estimate by a non parametric method. In one hand our estimator is a generalization of the classical kernel estimator for continuous time processes with values in Euclidean space (presented for example in Bosq, D. (1998)) to the Riemannian submanifold. In the other hand, our proposal is also a generalization of Pelletier's estimator for continuous time processes. In this talk, we will give some asymptotical results on the behavior of the estimator under some mixing conditions.

**Keywords.** *Kernel Density estimation, Riemannian manifolds, Mixing condition, time continuous process.*

# 1 Introduction

The issue of kernel density estimation has been treated by very large number of papers in the literature when the dataset lives in Euclidean space. But, in many situations the observations belong to a Riemannian manifold, $\mathcal{M}$, as for example in biology (we refer to Mardia and al. (2008) and reference therein), spatial statistics, geology, image analysis (as shown, for example in Pennec, X. (2006). The Riemannian manifold $\mathcal{M}$ can be either unknown (we refer, for example, Aamari, E. and Levrard, C. (2019) or Berenfeld, C. and Hoffmann, M. (2021) or Khardani, S. and Yao, A. F. (2022) and references therein) or known. In this work, we focus in the last case. Namely we are interested in the kernel estimation of the density based on observations on $\mathcal{M}$. The literature in this setting is very limited. When the observations are i.i.d. the interesting work of Pelletier, B. (2005) is the reference. But, we can also refer to Kim, Y. and Park, H. (2013), Henry and al. (2013), Berry, T. and Sauer, T. (2017), Kerkyacharian and al. (2020) and Berenfeld, C. and Hoffmann, M. (2021) and references therein.

This work deals with the estimation of the marginal density of stationary time continuous process $(X_t, t \in \mathbb{R})$, where the $X_t$'s are distributed as a random variable $X$ with values in a subriemannian manifold, $(\mathcal{M}, g)$, of $\mathbb{R}^d (2 \leq d < \infty)$ endowed with a measure $\mu_g$ and $(\mathcal{M}, d_g)$ is complete, where $d_g$ is the metric induced by $g$. Let $f$ be the density of $X$ with respect to $\mu_g$. We propose here a kernel estimator of $f$. In one hand our estimator is a generalization (to the Riemannian manifold setting) of the classical kernel estimator for continuous time processes (presented for example in Bosq, D. (1998)) with values in Euclidean space. In the other hand, our proposal is also a generalization of Pelletier's estimator for continuous time processes.

In this talk, we will give some asymptotical results on the behavior of the estimator under some mixing conditions.

# 2 General framework

This work concerns any mesurable continuous time stationary process, $(X_t, t \in \mathbb{R})$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values on Riemannian submanifold $(\mathcal{M}, g)$ of $\mathbb{R}^d$ $(2 \leq d < \infty)$ such that the $X_t$'s are distributed as a random variable $X$ with unknown density $f$ on $\mathcal{M}$.

We assume that $(\mathcal{M}, g)$ is endowed with a measure, $\mu_g$ and is geodesically complete compact without boundary. Consequently, $(\mathcal{M}, d_g)$ is a complete metric space, where $d_g$ is the metric induced by $g$. From now, to ease the reading we set $\mu := \mu_g$ and $d(.,.) = d_g(.,.)$. For all

$x \in \mathcal{M}$ we denote by $T_x\mathcal{M}$, the tangent space to $\mathcal{M}$ at $x$, $\exp_x : T_x\mathcal{M} \to \mathcal{M}$, the exponential map at $x$, $exp_x^{-1}$ its inverse and $0_x = exp_x^{-1}(x)$. Let $|g_x(v)|^{1/2} = \frac{d\mu}{d\mu_x}(\exp_x(v)) = \frac{d\mu_{\exp_x^* g}}{d\mu_x}(v)$ is the density of $\mu_{\exp_x^* g}$ with respect to $\mu_x$, the Lebesgue measure on $T_x(\mathcal{M})$.

We assume that the injectivity radius of $\mathcal{M}$ is such that, $inj(\mathcal{M}) > 0$ and consider only regular balls in $\mathcal{M}$. We recall that a regular (or convex) ball, $B(x,h)$, is such that $h < h^*$ where $h^* = \min\{inj(\mathcal{M}), \frac{\pi}{2\sqrt{\kappa}}\}$ with $\kappa$ the supremum of sectional curvatures of $\mathcal{M}$ (see for example [**?**] for definition) if this upper bound is positive, and $\kappa = 0$ otherwise. Then, $B(x,h) = \exp_x B(h)$, where $B(h)$ is the ball centred at $x$ with radius $h$.

Our aim is to estimate the unknown density $f$ on $\mathcal{M}$ from the data $(X_t, 0 \leq t \leq T)$ with $T > 0$.

## 2.1 Our kernel density estimator

We propose the kernel estimator of $f$ such that for all $x \in \mathcal{M}$:

$$f_T(x) = \frac{1}{T}\int_0^T \frac{1}{h_T^d} \frac{K\left(\frac{d(x,X_t)}{h_T}\right)}{|g_x(\exp_x^{-1}(X_t))|^{1/2}} dt, \tag{1}$$

where the kernel $K$ is a function : $\mathbb{R}_+ \longrightarrow \mathbb{R}$ and the bandwidth $h_T < h^*$ and $\lim_{T\to\infty} h_T = 0(+)$.

We study the asymptotic behavior of $f_T$ under the following assumptions.

### Assumptions

The dependence of the process will be measured by means of $\alpha$-mixing. Then, we consider the $\alpha$-mixing coefficients of the process $(X_t, t \in \mathbb{R})$ is defined by :

$$\alpha(u) = \sup_{t\in\mathbb{R}} \sup_{A\in\sigma(X_s, s\leq t),\, B\in\sigma(X_s, s\geq t+u)} \{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|\}, u \geq 0.$$

**HK:**

1. $K$ satisfies a Lipschitz condition,
2. *supp* $K = [0;1]$,
3. $\int K(\|v\|)dv = 1$,
4. $\int vK(\|v\|)dv = 0_x$ or at least the vector, $\int vK(\|v\|)dv$, is orthogonal to span$\{\text{grad} f(x)\}$
5. $\int \|v\|^2 K(\|v\|)dv < +\infty$.

**$\mathbf{Hh_T}$** : $Th_T^d \longrightarrow +\infty$ , $h_T \longrightarrow 0$ and $h_T < \frac{h^*}{2}$, when $T \to +\infty$

Assume that $h_T < \frac{h^*}{2}$, to ensure that $x$ is locally a central point when using the kernel estimator

**Hf :** The probability density $f(.)$ is bounded, twice continuously differentiable at any $x \in \mathcal{M}$

**H($\Gamma$,q) :** The exists $\Gamma \in \mathcal{B}(\mathbb{R}^2)$ containing $D = \{(s,t) \in \mathbb{R}^2 : s = t\}$ and $q \in ]2; +\infty]$ such that :

**(1)** $\tilde{f}_{s,t}$ and $\tilde{f}_{s,t} * exp_x$ exists for $(s,t) \notin \Gamma$ ;

**(2)** $\delta_{x,q}(\Gamma) = \sup_{(s,t)\notin\Gamma} \|\tilde{f}_{s,t} * exp_x\|_{L^q(\mathbb{R}^{2d})} < +\infty$ ;

**(3)** there exists a positive constant $l_\Gamma$ such that

$$l_\Gamma = \limsup_{T\to+\infty} \frac{1}{T} \int_{[0,T]^2\cap\Gamma} dsdt,$$

where $f_{(X_s,X_t)}$ denote the joint probability density of $(X_t)$ and

$$\tilde{f}_{s,t} = f_{(X_s,X_t)} - f \otimes f, \quad \forall (s,t) \in [0,T]^2$$
$$\tilde{f}_{s,t} * exp_x := \tilde{f}_{s,t}\left(exp_x(\cdot), exp_x(\cdot)\right), \quad \forall (s,t) \in [0,T]^2, \ \forall x \in \mathcal{M}.$$

**M($\gamma, \beta$) :** The mixing coefficients of the process $(X_t)$ satisfies $\alpha(|s-t|) = \gamma|s-t|^{-\beta}$ for $(s,t) \notin \Gamma$, where $\gamma > 0$ and $\beta > 0$.

# 3 Consistency results

## 3.1 Rates of convergence in Mean Squared Error(MSE)

**Theorem 3.1.** Under assumptions **HK**, **$\mathbf{Hh_T}$**, **Hf**, **H($\mathbf{\Gamma}$, q)** and **M**$(\gamma, \beta)$ hold for $\Gamma, q, \gamma$ and $\beta \geq \max(\frac{2(q-1)}{q-2}, 1)$.

$$MSE(x) := \mathbb{E}\left(f_T(x) - f(x)\right)^2 = O\left(h_T^4 + \frac{1}{Th_T^d}\right) ; \tag{2}$$

Furthermore since $f$ and $K$ are bounded, and $\mathcal{M}$ is compact, then

$$\sup_{x\in\mathcal{M}} MSE(x) = O\left(h_T^4 + \frac{1}{Th_T^d}\right). \tag{3}$$

## 3.2 Rates of convergence in Mean Integrated Squared Error(MISE)

**Theorem 3.2.** Under assumptions **HK**, $\mathbf{Hh_T}$, **Hf**, $\mathbf{H(\Gamma, q)}$ and $\mathbf{M}(\gamma, \beta)$ hold for $\Gamma, q, \gamma$ and $\beta \geq \max(\frac{2(q-1)}{q-2}, 1)$.

Since $f$ and $K$ are bounded, and $\mathcal{M}$ is compact, we have

$$MISE := \int_{\mathcal{M}} MSE(x)dx = O\left(h_T^4 + \frac{1}{Th_T^d}\right). \tag{4}$$

## 3.3 Convergence in probability

**Theorem 3.3.** Under the assumptions **HK**, **Hf**, $\mathbf{Hh_T}$, $\mathbf{H(\Gamma, q)}$ and $\mathbf{M}(\gamma, \beta)$ hold for $\Gamma, q, \gamma$ and $\beta \geq \max\left(\frac{2(q-1)}{q-2}, 1\right)$.

If $Th_T^{4+d}(\log T)^{-1} \to 0$ and $Th_T^{\frac{d(\beta+1)}{\beta-1}}(\log T)^{-1} \to \infty$ , then

$$f_T(x) - f(x) = \mathcal{O}_p\left(\left(\frac{\log T}{Th_T^d}\right)^{1/2}\right). \tag{5}$$

## 3.4 Almost sure convergence

Now to get the almost sure consistency rate result, we apply the Borel-Cantelli Lemma 4.2 of Bosq, D. (1998) to the sample paths of $Q_T = \left(\frac{\log T}{Th_T^d}\right)^{1/2}(f_T(x) - \mathbb{E}(f_T(x)))$.

Assume that the paths of $Q_T = \left(\frac{\log T}{Th_T^d}\right)^{1/2}(f_T(x) - \mathbb{E}(f_T(x)))$ are uniformly continuous with probability 1 and the bandwidth $h_T$ is such that :

$$\varphi_\eta(T) = C\left(T^{-1}h_T^{-\frac{d(\beta+1)}{\beta-1}}\log T\right)^{\frac{\beta-1}{2}} \text{ is decreasing} \tag{6}$$

$$Th_T^{\frac{d(\beta+1)}{\beta-3}}(\log T)^{-\frac{\beta-1}{\beta-3}}g(T)^{\frac{-2}{\beta-3}} \to \infty \tag{7}$$

where $g(T) = \log(T)(\log(\log T))^{1+\epsilon}$, and $\epsilon > 0$ is an arbitrary small real number. Then we get the following strong consistency result.

**Theorem 3.4.** Under the assumptions (6), (7), **HK**, **Hf**, $\mathbf{Hh_T}$, $\mathbf{H(\Gamma, q)}$ and $\mathbf{M}(\gamma, \beta)$ hold for $\Gamma, q, \gamma$ and $\beta \geq \max\left(\frac{2(q-1)}{q-2}, 3\right)$. If $Th_T^{4+d}(\log T)^{-1} \to 0$ , then

$$f_T(x) - f(x) = \mathcal{O}_{a.s.}\left(\left(\frac{\log T}{Th_T^d}\right)^{1/2}\right). \tag{8}$$

**Acknowledgments**

This work was supported by the Data Science Institute program in Côte d'Ivoire of Ecole polytechnique Paris, l'X (France) and funding from Félix Houphouët Boigny University and Lagunes University, Abidjan (Côte d'Ivoire) and University of Clermont Auvergne (France).

# References

Aamari, E. and Levrard, C. (2019), Non-asymptotic rates for manifold, tangent space and-curvature estimation, *The Annals of Statistics*, 47(1):177 – 204.

Berenfeld, C. and Hoffmann, M. (2021), Density estimation on an unknown submanifold, *Electronic Journal of Statistics*, 15(1):2179 – 2223

Berry, T. and Sauer, T. (2017), Density estimation on manifolds with boundary, *Computational Statistics & Data Analysis*, 107:1–17.

Bosq, D. (1998), *Nonparametric statistics for stochastic processes , estimation and prediction.Second Edition*, Springer.

Cleanthous, G., Georgiadis, A. G., Kerkyacharian, G., Petrushev, P., and Picard, D. (2020), Kernel and wavelet density estimators on manifolds and more general metric spaces, *Bernoulli*, 26(3):1832 – 1862.

Gallot, S., Hulin, D., and Lafontaine, J. (2004), Riemmanian geometry .

Henry, G., Munoz, A., and Rodriguez, D. (2013), Locally adaptive density estimation on riemannian manifolds, *SORT-Statistics and Operations Research Transactions*, pages 111–130.

Khardani, S. and Yao, A. F. (2022), Nonparametric recursive regression estimation on riemannian manifolds, *Statistics & Probability Letters*, 182:109274.

Kim, Y. and Park, H. (2013), Geometric structures arising from kernel density estimation on riemannian manifolds, *Journal of Multivariate Analysis*, 114:112–126.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008), A multivariate von misesdistribution with applications to bioinformatics, *Canadian journal of statistics*, 36(1):99–109.

Pelletier, B. (2005), Kernel density estimation on riemannian manifolds, *Statistics & probability letters*, 73(3):297–304.

Pennec, X. (2006), Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements, *Journal of mathematical imaging and vision*, 25(1):127–154.

# Index des auteurs

# Merci

## à nos soutiens

JdS 2024
55e journées de statistique de la SFdS

suez le lyre

Ceva Santé Animale

HORIANA

SLPV analytics

aixial GROUP AN ALTEN COMPANY

Betclic

casden BANQUE POPULAIRE

DIGITAL AQUITAINE

RÉGION Nouvelle-Aquitaine

université de BORDEAUX

BORDEAUX INP

MARGAUx Fédération Mathématique de Recherche en Région Nouvelle-Aquitaine

cea

Réseau de Recherche Impulsion PHDS | Public Health Data Science Bordeaux Network / université de BORDEAUX

Digital Public Health Graduate Program / université de BORDEAUX

Département de recherche Santé publique / université de BORDEAUX

cnrs

Inria

Inserm