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-β 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 θ 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-β42 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.