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.