funHDDC : extension du package R pour le clustering de courbes fonctionnelles multivariées

A. Schmutz[a,b,d,e], **J. Jacques**[b], **C. Bouveyron**[c], **L. Chèze**[d], **P. Martin**[a,e]

[a]Lim France, Chemin Fontaine de Fanny, Nontron, France
aschmutz@lim-group.com & pmartin@lim-group.com
[b]Université de Lyon, Lyon 2, ERIC EA3083, Lyon, France
julien.jacques@univ-lyon2.fr
[c]Université Côté d'Azur, LJAD & Asclepios, Nice, France
charles.bouveyron@math.cnrs.fr
[d]Université de Lyon, Lyon 1, LBMC UMR T9406, Lyon, France
laurence.cheze@univ-lyon1.fr
[e]CWD-Vetlab, Ecole Nationale Vétérinaire d'Alfort, F-94700, France

Mots clefs : Clustering, Données Fonctionnelles, Package

D'après une étude du cabinet Gartner datant de 2017, ils prévoyaient que 8.3 milliard d'objets connectés seraient vendus dans le monde en 2017 et 20.5 milliards d'ici 2020. L'essor des objets connectés pour accompagner notre quotidien facilite la collecte de données à haute fréquence. Ce type de données peut être classé dans la catégorie des données fonctionnelles : une variable quantitative qui évolue au cours du temps. Ainsi une donnée fonctionnelle univariée X est représentée par une unique courbe, $X(t) \in \mathbb{R}$, $\forall t \in [0,T]$. Dans le cas multivarié on peut écrire $\boldsymbol{X} = \boldsymbol{X}(t)_{t\in[0,T]}$ avec $\boldsymbol{X}(t) = (X^1(t),...,X^p(t))' \in \mathbb{R}^p$, $p \geq 2$. Avec la déclinaison des objets connectés à tous les domaines, les besoins en méthodes pour faciliter la modélisation et la compréhension des données fonctionnelles multivariées vont augmenter.
De nombreuses méthodes existent pour le clustering de données fonctionnelles univariées (James et Sugar (2003), Tarpey et Kinateder (2003), Chiou et Li (2007), Bouveyron et Jacques (2011), Jacques et Preda (2013), Bouveyron et al. (2015)), et la plupart sont disponibles sous la forme de packages R. En revanche, les méthodes pour le cas des données fonctionnelles multivariées sont nombreuses dans la littérature (Singhal et Seborg (2005), Tokushige et al. (2007), Kayano et al. (2010), Ieva et al. (2013), Yamamoto et Hwang (2017)) mais seule l'une d'entre elles est disponible sous la forme d'un package R (Jacques et Preda, 2014). Cette dernière présente des limites : une partie de l'information disponible est ignorée car seules les premières composantes principales sont modélisées. De plus, cette utilisation d'un nombre limité de composantes entraîne des problèmes d'inférence.
Nous proposons un modèle de clustering qui étend les travaux de Jacques et Preda (2013) en modélisant toutes les composantes principales de variance non-nulle. Notre méthode de clustering nécessite le lissage des données dans une base de fonctions, les données sont ensuite projetées dans les sous-espaces spécifiques aux groupes à l'aide d'une analyse en composante principale fonctionnelle. Les scores ainsi obtenus sont considérés comme des variables aléatoires dont la distribution de probabilité est spécifique du groupe, un modèle de mélange gaussien est alors utilisé sur les scores. Etant dans le cas non supervisé, un algorithme de type EM est mis en place pour faire ces calculs. Il réalise une première étape (E) d'estimation de la probabilité d'appartenance de la courbe au groupe, puis dans une seconde étape (M), une analyse en composantes principales fonctionnelle par groupe est appliquée, permettant de mettre à jour les paramètres du modèle. Ces 2 étapes E et M sont itérées jusqu'à convergence de la log-vraisemblance. L'efficacité de ce modèle a été prouvée sur données simulées et sur données

réelles (Schmutz et al., 2017). Nous présentons ici le package R associé à cette fonction en s'appuyant sur plusieurs exemples applicatifs.

Références

[1] James, G., Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98**(462), 397-408.

[2] Tarpey, T., Kinateder, K. (2003). Clustering functional data. *Journal of Classification*, **20**(1),93-114.

[3] Chiou, J.M., Li, P.L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society Series B Statistical Methodology*, **69**(4), 679-99.

[4] Bouveyron, C., Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, **5**(4), 281-300.

[5] Jacques, J., Preda, C. (2013). Funclust: a curves clustering method using functional random variable density approximation. *Neurocomputing*, **112**, 164-71.

[6] Bouveyron, C., Come, E., Jacques, J. (2015). The discriminative functional mixture model for the analysis of bike sharing systems. *Annals of Applied Statistics*, **9**(4), 1726-60.

[7] Singhal, A., Seborg, D. (2005). Clustering multivariate time-series data. *Journal of Chemometrics*, **19**, 427-38.

[8] Tokushige, S., Yadohisa, H., Inada, K. (2007). Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, **22**, 1-16.

[9] Kayano, M. Dozono, K., Konishi, S. (2010). Functional Cluster Analysis via Orthonormalized Gaussian Basis Expansions and Its Application. *Journal of Classification*,**27**, 211-30.

[10] Ieva, F., Paganoni, A., Pigoli, D., Vitelli, V. (2013). Multivariate Functional Clustering for the Morphological Analysis of ECG Curves. *Journal of the Royal Statistical Society, Series C Applied Statistics*, **62**(3), 401-18.

[11] Yamamoto, M., Hwang, H. (2017). Dimension-Reduced Clustering of Functional Data via Subspace Separation. *Journal of Classification*, **34**, 294-326.

[12] Schmutz, A., Jacques, J., Bouveyron, C., Chèze, L., Martin, P. (2017). Clustering multivariate functional data in group-specific functional subspaces. *Preprint HAL n ° 01652467.*