Données fonctionnelles multivariées issues d'objets connectés : une méthode pour classer les individus

Amandine Schmutz, Julien Jacques, Charles Bouveyron, Laurence Cheze, Pauline Martin

Données fonctionnelles multivariées issues d'objets connectés : une méthode pour classer les individus

Amandine Schmutz [1,2,4], Julien Jacques [2], Charles Bouveyron [3], Laurence Chèze [4] & Pauline Martin [1]

[1] *Lim France, Chemin Fontaine de Fanny, 24300 Nontron, France et aschmutz@lim-group.com & pmartin@lim-group.com*

[2] *Université de Lyon, Lyon 2, ERIC EA3083, Lyon, France et julien.jacques@univ-lyon2.fr*

[3] *Université Côte d'Azur, INRIA Sophia-Antipolis, Laboratoire J.A. Dieudonné, UMR CNRS 7351 & Equipe Epione, Nice, France et charles.bouveyron@math.cnrs.fr*

[4] *Université de Lyon, Lyon 1, LBMC UMR T9406, Lyon, France et laurence.cheze@univ-lyon1.fr*

Résumé. L'émergence des objets connectés pour tous les aspects de la vie quotidienne entraine des besoins croissants de méthodes pour analyser des données fonctionnelles multivariées. Ce travail propose une méthode de clustering (Schmutz *et al*, 2017) de façon à faciliter la modélisation et la compréhension de ces données fonctionnelles multivariées. Cette méthode est basée sur un modèle de mélange latent fonctionnel qui répartit les individus dans des sous-espaces fonctionnels spécifiques aux groupes à l'aide d'une analyse en composante principale multivariée fonctionnelle. Un algorithme de type EM est proposé pour l'inférence du modèle et le choix des hyper paramètres se fait par le biais de la sélection de modèle. L'efficacité du modèle sera testée sur un exemple original de prédiction de vitesse, pour des exemples classiques se reporter à Schmutz *et al* (2017).

Mots-clés. Apprentissage et classification, Analyse des données, fouille de données, Grande dimension, données massives, Statistique computationnelle, Données fonctionnelles

Abstract. The emergence of numerical sensors in many aspects of everyday life leads to an increasing need of methods to analyze multivariate functional data. This work presents a clustering technique (Schmutz *et al*, 2017) in order to ease the modeling and understanding of those multivariate functional data. This method is based on a functional latent mixture model which fits the data in group-specific functional subspaces through a multivariate functional principal component analysis. An EM-like algorithm is proposed for model inference and the choice of hyper-parameters is carried out through model selection. Model efficiency will be tested on an original example of speed prediction, for classical example see Schmutz *et al* (2017).

Keywords. Learning and classification, Data analysis, Big data, Computational statistics, Functional data

1 Introduction

Les nouvelles technologies facilitent la collecte de données à haute fréquence. Par exemple dans le domaine sportif, les athlètes portent des dispositifs permettant une collecte constante de données pendant leur entrainement afin d'améliorer leur performance ou de prévenir des blessures grâce au suivi de leurs constantes physiques. Ce type de données peut être catégorisé comme des données fonctionnelles : des valeurs quantitatives qui évoluent au cours du temps. Dans le cas univarié, une donnée fonctionnelle X est représentée par une unique courbe, $X(t) \in \mathbb{R}, \forall t \in [0,T]$. Avec la croissance du marché des objets connectés de plus en plus de données sont collectées pour un même individu. Un individu est alors représenté par plusieurs courbes. On peut alors écrire : $\boldsymbol{X} = \boldsymbol{X}(t)_{t\in[0,T]}$ avec $\boldsymbol{X}(t) = (X^1(t), ..., X^p(t))' \in \mathbb{R}^p, p \geq 2$. Des exemples univariés et bivariés sont disponibles dans Ramsey et Silverman (2005).

De nombreux travaux 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), Bouveyron *et al.* (2018). Mais beaucoup moins de méthodes existent pour le cas multivarié avec des techniques de réduction de la dimension : Yamamoto (2012), Yamamoto et Terada (2014), Jacques et Preda (2014) et Yamamoto et Hwang (2017). L'objectif de ce travail est de proposer une méthode de clustering fonctionnel multivariée qui est moins restrictive que la méthode proposée par Jacques et Preda (2014).

2 Modèle

2.1 Reconstruction des données fonctionnelles

En pratique l'expression fonctionnelle des courbes observées est inconnue et nous avons uniquement accès aux observations discrètes mesurées à des temps précis. Une technique classique pour reconstruire la forme fonctionnelle est d'exprimer les courbes dans un espace de dimension fini engendré par une base de fonctions. Soit $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ un ensemble de n courbes p-variées avec $\boldsymbol{X}_i = (X_i^1, ..., X_i^p)$. On suppose que chaque courbe est décrite par une combinaison linéaire de bases de fonctions :

$$X_i^j(t) = \sum_{r=1}^{R_j} c_{ir}^j(X_i^j)\phi_r^j(t),$$

avec $i \in \{1, ..., n\}, j \in \{1, ..., p\}, R_j$ le nombre de bases de fonctions et $(\phi_r^j(t))_{1\leq r\leq R_j}$ la base de fonction pour la *j-ième* composante de la courbe multivariée. Les coefficients c_{ir}^j peuvent ensuite être concaténés au sein d'une matrice :

$$\mathbf{C} = \begin{pmatrix} c_{11}^1 & ... & c_{1R_1}^1 & c_{11}^2 & ... & c_{1R_2}^2 & ... & c_{11}^p & ... & c_{1R_p}^p \\ & & & & ... & & & & & \\ c_{n1}^1 & ... & c_{nR_1}^1 & c_{n1}^2 & ... & c_{nR_2}^2 & ... & c_{n1}^p & ... & c_{nR_p}^p \end{pmatrix}.$$

Et les matrices de bases de fonctions $(\phi_r^j)_{1\leq r\leq R_j}$ peuvent être concaténées au sein de $\boldsymbol{\phi}(t)$ tel que :

$$\boldsymbol{\phi(t)} = \begin{pmatrix} \phi_1^1(t) & ... & \phi_{R_1}^1(t) & 0 & ... & 0 & 0 & ... & 0 \\ 0 & ... & 0 & \phi_1^2(t) & ... & \phi_{R_2}^2(t) & 0 & ... & 0 \\ & & & & ... & & & & \\ 0 & ... & 0 & 0 & ... & 0 & \phi_1^p(t) & ... & \phi_{R_p}^p(t) \end{pmatrix}.$$

Ansi avec ces notation on peut alors écrire : $\boldsymbol{X}(t) = \boldsymbol{C}\boldsymbol{\phi}'(t)$.

2.2 Un nouveau modèle de clustering pour données fonctionnelles multivariées

Notre objectif est de séparer $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ en K clusters. Soit Z_{ik} la variable latente tel que $Z_{ik} = 1$ si $\boldsymbol{X}_i$ appartient au cluster k, 0 sinon. Afin de faciliter la présentation du modèle, supposons dans un premier temps que les valeurs z_{ik} de Z_{ik} sont connues pour tout $1 \leq i \leq n$ et $1 \leq k \leq K$. (Notre but en pratique est de les estimer à partir des données). Ainsi, $n_k = \sum_{i=1}^n z_{ik}$ correspond au nombre de courbes du cluster k.

Supposons que ces courbes peuvent être décrites dans un sous-espace fonctionnel de plus faible dimension $d_k \leq R$, $k = 1, ..., K$. La base spécifique au groupe est obtenue à partir de $\{\phi_r^j\}_{(1\leq j\leq p),(1\leq r\leq R_j)}$ par transformation linéaire à l'aide d'une analyse en composante principale fonctionnelle : $\varphi_{kj}(t) = \sum_{l=1}^R q_{k,jl}\phi_l(t)$ avec $(q_{k,jl})$ les coefficients des fonctions propres exprimés dans la base initiale ϕ.

Soit $(\boldsymbol{\delta}_i^k)_{1\leq i\leq n_k}$ les scores d'analyse en composante principale fonctionnelle des n_k courbes du cluster k. On assume que ces scores suivent une distribution Gaussienne tels que $\boldsymbol{\delta}_i^k \sim \mathbb{N}(\boldsymbol{\mu}_k, \boldsymbol{\Delta}_k)$ avec $\boldsymbol{\mu}_k \in \mathbb{R}^R$ la fonction moyenne, et $\boldsymbol{\Delta}_k$ pouvant s'écrire :

$$\boldsymbol{\Delta}_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kd_k} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & b_k \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} R - d_k \end{matrix}$$

Cette hypothèse sur $\boldsymbol{\Delta_k}$ permet de modéliser finement la variance des premières d_k composantes principales, les dernières étant considérées comme du bruit et sont alors modélisées par un unique paramètre b_k.

En pratique les z_{ik} sont inconnus et notre but est de les prédire. Pour cela, un algorithme de type EM est proposé pour estimer les paramètres du modèle. Les z_{ik} seront ensuite estimés par Maximum a posteriori.

2.3 Estimation du modèle

En classification non supervisée, l'estimation des paramètres d'un modèle de mélange est traditionnellement réalisée en maximisant la vraisemblance à l'aide d'un algorithme EM. Le principe de cet algorithme est d'alterner entre une phase dite *Expectation*, qui calcule l'espérance de la log-vraisemblance complétée ; et une phase dite *Maximisation*, qui calcule les paramètres maximisant l'espérance de la log-vraisemblance complétée trouvée à l'étape précédente.

La log-vraisemblance complétée des courbes observées selon le modèle décrit précédement est la suivante :

$$\begin{aligned}\ell_c(\theta) &= -\frac{1}{2}\sum_{k=1}^{K} n_k \left[\sum_{j=1}^{d_k}\left(\log(a_{kj}) + \frac{q_{kj}^t C_k q_{kj}}{a_{kj}}\right)\right. \\ &\quad \left. + \sum_{j=d_k+1}^{R}\left(\log(b_k) + \frac{q_{kj}^t C_k q_{kj}}{b_k}\right) - 2\log(\pi_k)\right] + \frac{R}{2} log(2\pi),\end{aligned}$$

où $\theta = (\pi_k, \mu_k, a_{kj}, b_k, q_{kj})_{kj}$ pour $1 \leq j \leq d_k$ et $1 \leq k \leq K$, q_{kj} est la j-ième colonne de la matrice des coefficients des fonctions propres du groupe k et $C_k = \frac{1}{n_k}\sum_{i=1}^{n} Z_{ik}(c_i - \mu_k)^t(c_i - \mu_k)$. Comme l'appartenance aux groupes Z_{ik} est inconnue, il est nécessaire de calculer leur espérance conditionnelle (*E step*) avant de maximiser l'espérance de la vraisemblance complétée (*M step*).

L'étape M nécessite la réalisation d'une ACP fonctionnelle par classe où chaque donnée fonctionnelle est pondérée par sa probabilité a posteriori d'appartenance à la classe.

3 Application sur des données issues d'un objet connecté

L'objectif de cette étude est de trouver la combinaison de modèles permettant d'obtenir la prédiction de vitesse la plus fine. La performance de notre modèle de clustering va ainsi être évaluée sur cette base.

3.1 Les données

Afin de développer un outil d'aide à l'entrainement pour les cavaliers, Lim France a positionné un accéléromètre et un gyroscope sur le garrot de chevaux de façon à récolter des données à chaque foulée selon les 3 axes du mouvement (x, y, z). Une base de données a été constituée en couplant ces mesures à celles d'une méthode de référence : l'acquisition

de vidéos rapides. Cette dernière permet de mesurer la vitesse réelle à chaque foulée dans un champ d'étude de 26 mètres. A ce jour, 2685 foulées de galop en ligne droite ont été mesurées sur 44 chevaux équipés du dispositif. L'objectif de cette étude est de prédire la vitesse instantanée du cheval à partir des données d'accéléromètre et de gyroscope avec une précision de 0.6 m/s.

3.2 Résultats

Le logiciel R 3.3.2 a été utilisé pour les analyses. Un jeu de données d'entrainement et un jeu de données de test sont constitués par tirage aléatoire de la base de données avec le ratio 80/20. Les données ont été scindées en 2 sous-groupes homogènes à l'aide de notre méthode de clustering fonctionnel multivarié. Puis un modèle de régression non paramétrique fonctionnelle (Ferraty et Vieu, 2009) est utilisé pour chaque sous-groupe. Les résultats d'une simulation sont présentés en Figure 1.

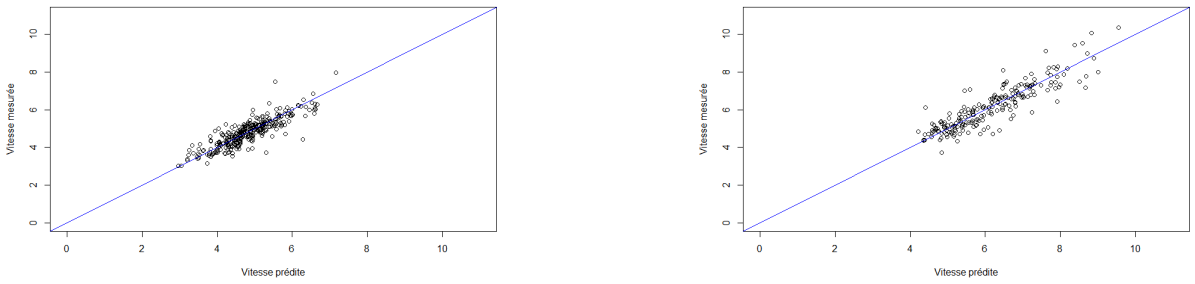

FIGURE 1 – Vitesse prédite versus vitesse mesurée pour les données de test sur une simulation : résultats pour le groupe 1 (à gauche) et le groupe 2 (à droite)

On peut voir dans le cas du 2ème groupe que le nuage de points est moins resserré autour de la première bissectrice que le 1er groupe, mais l'erreur de prédiction reste faible. Avec ce modèle on obtient une moyenne de 13% d'erreur de prédiction supérieure à 0.6 m/s.

4 Conclusion

Pour conclure les résultats obtenus suite à la succession de notre algorithme de clustering et une régression non paramétrique fonctionnelle multivariée surpassent les modèles biomécaniques et physiques existants pour le cheval. En effet l'intégration de l'accélération pour calculer la vitesse est la technique la plus exploitée actuellement, cependant cette méthode réalise une intégration des erreurs entrainant une perte de précision importante. L'utilisation de modèles d'analyse de données fonctionnelles permettent d'atteindre les attentes de précision des cavaliers et des entraineurs, celle-ci étant un critère indispensable

à l'utilisation d'objets connectés par les professionnels. Ces modèles permettent aussi de proposer un paramètre de la locomotion du cheval, la vitesse instantanée, qui n'est fournie par aucun autre objet connecté à ce jour.

Bibliographie

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

[2] Ramsay, J.O. et Silverman B.W. (2005), *Functional data analysis 2nd Edition*, Springer Series in Statistics, New-York.

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

[4] Tarpey, T. et Kinateder, K. (2003), Clustering functional data, *Journal of Classification*, 20, 93–114.

[5] Chiou, J.M. et Li, P.L. (2007), Functional clustering and identifying substructures of longitudinal data, *Journal of the Royal Statistical Society Series B Statistical methodology*, 69, 679–699.

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

[7] Jacques, J. et Preda, C. (2013), Funclust : a curves clustering method using functional random variable density approximation, *Neurocomputing*, 112, 164–171.

[8] Bouveyron C., Come, E. et Jacques, J. (2015), The discriminative functional mixture model for the analysis of bike sharing systems. *Annals of Applied Statistics*, 9, 1726–1760.

[9] Bouveyron, C., Bozzi, L., Jacques, J. et Jollois, F.-X. (2018), The Functional Latent Block Model for the Co-Clustering of Electricity Comsumption Curves, Journal of the Royal Statistical Society, Series C, in press.

[10] Yamamoto, M. (2012), Clustering of Functional Data in a Low-Dimensional Subspace, *Advances in Data Analysis and Classification*, 6, 219–247

[11] Jacques, J. et Preda, C. (2014), Model based clustering for multivariate functional data, *Computational Statistics and Data Analysis*, 112, 164–171.

[12] Yamamoto, M., Terada, Y. (2014), Functional Factorial k-Means Analysis. *Computational Statistics and Data Analysis*, 79, 133–148.

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

[14] Ferraty, F., Vieu, P. (2009), Additive prediction and boosting for functional data. *Computational Statistics and Data Analysis*, 53, 1400–1413.