Clustering et détection d'anomalies dans les données fonctionnelles

Martial Amovin-Assagba, Irène Gannaz, Julien Jacques

Clustering et détection d'anomalies dans les données fonctionnelles

Martial AMOVIN-ASSAGBA[1,2], Irène GANNAZ[3] & Julien JACQUES[1].

[1] *Univ Lyon, Univ Lyon 2, ERIC EA3083, Lyon, France, julien.jacques@univ-lyon2.fr*
[2] *Arpege Master K, Saint-Priest, France, martial.amovin@masterk.com.*
[3] *Univ Lyon, INSA de Lyon, CNRS UMR 5208, Institut Camille Jordan, F-69621 Villeurbanne, France, irene.gannaz@insa-lyon.fr*

Résumé. Ce travail propose une méthode permettant simultanément de grouper en clusters et de détecter des anomalies dans des données fonctionnelles multivariées. Les données fonctionnelles sont décomposées dans une base de fonctions de dimension finie. La méthode repose ensuite sur des modèles de mélanges gaussiens contaminés parcimonieux sur cette décomposition. Un algorithme ECM est utilisé pour l'inférence du modèle. La performance du modèle est illustrée sur des données simulées.

Mots-clés. Détection d'anomalies, clustering, données fonctionnelles contaminées.

Abstract. This work proposes a method to group simultaneously into clusters and detect anomalies in multivariate functional data. Functional data are decomposed into a finite-dimensional basis functions. The method is then based on parcimonious contaminated Gaussian mixtures models on this decomposition. An ECM algorithm is used for model inference. The performance of the model is illustrated on simulated data.

Keywords. Anomaly detection, clustering, contaminated functional data.

1 Introduction

Les nouvelles méthodes d'acquisition de données favorisent la collecte massive des données avec une fréquence très élevée. Une manière de gérer ces amas de données est de les considérer comme des données fonctionnelles (Ramsay et Silverman, 2005), les observations étant assimilées à des courbes. Un des défis majeurs dans l'analyse de données fonctionnelles est la détection d'anomalies. Un certain nombre de travaux ont été proposés pour détecter des anomalies dans les données fonctionnelles univariées et multivariées. Certains travaux sont basés sur des approches de type kmeans robustes (Garcia-Escudero et Gordaliza, 2005), sur des fonctions de profondeur (Hubert et al. 2017), ou sur des techniques d'isolement (Staerman et al. 2019). Un point commun de ces méthodes est qu'elles requièrent toutes une connaissance a priori de la proportion des données anormales.

La détection de données anormales peut être grandement affectée par la présence de plusieurs clusters distincts au sein des données. Une approche intéressante est alors de

réaliser simultanément un clustering des données et une détection de données anormales. Dans cette optique, (Punzo et McNicholas, 2016) proposent, dans un cadre non fonctionnel, un modèle de mélange gaussien contaminé dans lequel chaque composante du mélange contient des données normales et/ou anormales.

Dans cet article, en s'inspirant de (Punzo et McNicholas, 2016) et sous l'hypothèse que les données fonctionnelles peuvent être décomposées dans une base de fonctions de dimension finie, nous proposons un modèle de mélange gaussien contaminé pour les données fonctionnelles multivariées. En plus de grouper les données en différents clusters, cette méthode permet de détecter des anomalies dans les données fonctionnelles. Dans les sections suivantes, nous présentons le modèle mathématique, son inférence, et des résultats obtenus sur simulations.

2 Modélisation

Soit $X = X(t), t \in [0,T]$, une variable aléatoire fonctionnelle de dimension p. Soient $X_i(t)_{\{1\leq i\leq n\}}$, n réplications indépendantes de la variable aléatoire fonctionnelle X et $x_i(t) = (x_i^1(t), x_i^2(t), \ldots, x_i^p(t))$, une réalisation de $X_i(t)$, $1 \leq i \leq n$. L'objectif est à la fois de grouper ces données en différents clusters et de détecter des données anormales.

2.1 Reconstruction fonctionnelle des données et description du modèle

En pratique, les expressions fonctionnelles des courbes observées ne sont pas connues et les données à disposition sont des observations discrètes enregistrées sur un nombre fini de points. Il est donc nécessaire de reconstruire la forme fonctionnelle des données. Pour ce faire, nous décomposons les courbes x_i^j dans une base de fonctions $\psi_m^j \in L_2, 1 \leq m \leq M_j$:

$$x_i^j(t) = \sum_{m=1}^{M_j} c_{im}^j \psi_m^j(t),$$

où M_j est le nombre de fonctions de base par composante et $\psi_m^j, 1 \leq m \leq M_j$ est la m-ème composante de la base de fonctions ψ^j. Le choix de la base de fonctions ainsi que du nombre de fonctions dépendent de la nature des données. Posons c_i le vecteur concaténant tous les coefficients de l'observation i obtenus pour chaque composante : $c_i = (c_i^1, \ldots, c_i^p)$ avec $c_i^j = (c_{i1}^j, \ldots, c_{iM_j}^j)$. Chaque donnée x_i peut être écrite sous la forme :

$$x_i(t) = \Psi(t) c_i^T \quad \text{avec} \quad \Psi(t) = \begin{pmatrix} \psi_1^1(t) & \cdots & \psi_{M_1}^1(t) & 0_{M_2} & \cdots & 0_{M_p} \\ 0_{M_1} & & & \psi_1^2(t) & \cdots & \psi_{M_2}^2(t) & \cdots & 0_{M_p} \\ \vdots & & & \vdots & \ddots & \vdots \\ 0_{M_1} & & & 0_{M_2} & \cdots & \psi_1^p(t) & \cdots & \psi_{M_p}^p(t) \end{pmatrix}.$$

Les modèles de mélanges gaussiens contaminés (Punzo et McNicholas, 2016) supposent que la densité de chaque cluster est un mélange de deux distributions gaussiennes multivariées dont l'une représente les observations normales et l'autre les données anormales. Cependant, de manière générale, la notion de densité de probabilité n'existe pas pour les variables aléatoires fonctionnelles. (Delaigle et Hall, 2010) ont montré que cette densité peut être approchée par la densité de probabilité des scores de l'Analyse en Composantes Principales Fonctionnelle (ACPF). Plusieurs travaux définissent alors des algorithmes de clustering probabilistes en considérant des modèles de mélange gaussiens sur ces scores (Schmutz et al. 2018). A noter que cette hypothèse est équivalente à supposer un modèle de mélange gaussien sur les coefficients c_i. De ce fait, la densité du mélange peut être écrite sous la forme :

$$h = \sum_{k=1}^{K} \pi_k \left[w_k f(c_i, \mu_k, \Sigma_k) + (1 - w_k) f(c_i, \mu_k, \eta_k \Sigma_k) \right] \tag{1}$$

où K est le nombre total de clusters, π_k est la probabilité d'appartenir au cluster k, w_k est la proportion de données normales du cluster k, μ_k et Σ_k sont respectivement la moyenne et la covariance du cluster k, η_k le facteur d'inflation de la variance pour les données anormales, et $f(\cdot, \mu_k, \Sigma_k)$ la densité gaussienne.

L'irrégularité des courbes observées peut entraîner l'utilisation de bases de fonctions ψ^j de relativement grande dimension, et ainsi une augmentation du nombre de paramètres à estimer, notamment du fait des matrices de covariance Σ_k. Pour contrer le fléau de la dimension, nous proposons d'utiliser la technique de réduction de dimension de (Schmutz et al. 2020), qui proposent de projeter les données de chaque cluster dans un sous-espace latent fonctionnel de faible dimension $d_k < M = \sum_{j=1}^{p} M_j$. À l'aide d'une ACPF, l'expression des fonctions propres de chaque cluster peut être obtenu à partir des fonctions ψ_m :

$$\varphi_{km}(t) = \sum_{l=1}^{M} s_{kml} \psi_l(t)$$

où s_{kml} sont les coefficients des fonctions propres. Les d_k premières composantes de la variance du cluster k, dans l'espace engendré par les fonctions propres, sont modélisées de manière fine par les d_k premières valeurs propres. Les $M - d_k$ dernières composantes sont modélisées de façon parcimonieuse par un seul paramètre b_k. La variance du cluster k dans l'espace propre s'écrit alors $\text{diag}(a_{k1}, \ldots, a_{kd_k}, b_k, \ldots, b_k)$.

2.2 Inférence du modèle

Un moyen usuel pour estimer les paramètres d'un modèle de mélange est l'algorithme EM. C'est un algorithme itératif qui alterne deux étapes: l'étape E (*Expectation*) qui calcule l'espérance de la log-vraisemblance complétée par les variables latentes, et l'étape M (*Maximization*) qui maximise cette espérance.

Deux variables latentes sont introduites dans le modèle. La première est la variable z_{ik}. $z_{ik} = 1$ si la donnée multivariée x_i appartient au cluster k et 0 sinon. La seconde variable introduite notée v_{ik} est en lien avec la présence de données anormales, $v_{ik} = 1$ si la donnée x_i du cluster k est une donnée normale et 0 sinon. L'expression de la log-vraisemblance complétée est alors :

$$\begin{aligned}\ell_c(\Phi) =& \sum_{i=1}^{n}\sum_{k=1}^{K}\{z_{ik}\log\pi_k + z_{ik}[v_{ik}\log w_k + (1-v_{ik})\log(1-w_k)]\} \\ =& -\frac{1}{2}\sum_{k=1}^{K}\left\{M(n_k - n_{k_n})\log\eta_k + n_k\left(\sum_{l=1}^{d_k}\log(a_{kl}) + \sum_{l=d_k+1}^{M}\log(b_k)\right)\right. \\ &\left. + n_k tr\left[\sum_{l=1}^{d_k}\frac{s'_{kl}W^{\frac{1}{2}}R_kW^{\frac{1}{2}}s_{kl}}{a_{kl}} + \sum_{l=d_k+1}^{B}\frac{s'_{kl}W^{\frac{1}{2}}R_kW^{\frac{1}{2}}s_{kl}}{b_k}\right]\right\} - \frac{nM}{2}\log(2\pi)\end{aligned}$$

où $n_k = \sum_{i=1}^{n} z_{ik}$ et $n_{k_n} = \sum_{i=1}^{n} z_{ik}v_{ik}$ sont respectivement le nombre de données et le nombre de données normales du cluster k, $R_k = \frac{1}{n_k}\left(\sum_{i=1}^{n} z_{ik}\left(v_{ik} + \frac{1-v_{ik}}{\eta_k}\right)(c_i - \mu_k)'(c_i - \mu_k)\right)$ est la matrice de covariance empirique du cluster k, W est la matrice du produit scalaire entre les fonctions de base Ψ et $\Phi = \{\Phi_1, \Phi_2\}$ avec $\Phi_1 = \{\pi_k, w_k, \mu_k, a_k, b_k, S_k\}_{k=1}^{K}$ et $\Phi_2 = \{\eta_k\}_{k=1}^{K}$. S_k est la matrice dont la $l-$ème colonne est s_{kl}.

En maximisant $\ell_c(\Phi)$, l'expression de l'estimation de Φ_2 est fonction de Φ_1 et vice versa. Nous proposons alors d'utiliser l'algorithme ECM (Expectation Conditional Maximization) où chaque étape M de EM est remplacée par deux étapes M. À la première étape M, nous fixons Φ_2, nous réalisons une ACPF par cluster et nous estimons les paramètres Φ_1; à la seconde étape M, nous fixons Φ_1 et nous estimons Φ_2.

3 Expériences

Dans cette section, nous présentons les données simulées et les résultats des expériences que nous avons réalisées sur ces données.

3.1 Données

Un jeu de données bivarié composé de 3 clusters a été simulé. Les proportions de mélange sont identiques $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$. Pour simplifier, nous considérons que pour chaque classe, la variance est commune pour toutes les dimensions intrinsèques d_k, c'est à dire $a_{kj} = a_k$ pour $k = 1, \ldots, K$ et $j = 1, \ldots, d_k$. L'objectif de ces expériences est d'une part d'évaluer l'efficacité de l'algorithme ECM en retrouvant les paramètres a_k, b_k et d_k, et d'autre part d'évaluer la performance de notre modèle du point de vue du clustering et de la détection d'anomalies. Nous fixons $a_1 = 150, a_2 = 100, a_3 = 50$, $b_1 = 5, b_2 =$

$3, b_3 = 10$ $\eta_1 = 30, \eta_2 = 50$ et $\eta_3 = 1$. Les dimensions intrinsèques de chaque cluster sont $d_1 = 5, d_2 = 20, d_3 = 10$. $\mu_1 = (1, 0, 50, 100, 0, \ldots, 0), \mu_2 = (0, 0, 80, 0, 40, 2, 0, \ldots, 0), \mu_3 = (0, \ldots 0, 20, 0, 80, 0, 0, 100)$. La proportion des données anormales est 5‰. La forme fonctionnelle des données est reconstruite à partir de 35 fonctions de Fourier. La Figure 1 présente les données simulées.

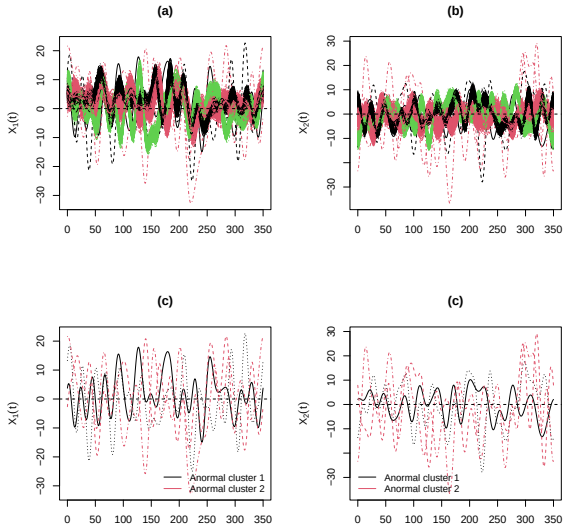

Figure 1: En haut : Données simulées. La première composante est à gauche et la seconde à droite. Chaque courbe est colorée selon son groupe. En bas : Courbes anormales. Les couleurs des courbes anormales diffèrent selon leur cluster d'attachement : noire pour les courbes anormales du groupe 1 et rouge pour celles du groupe 2.

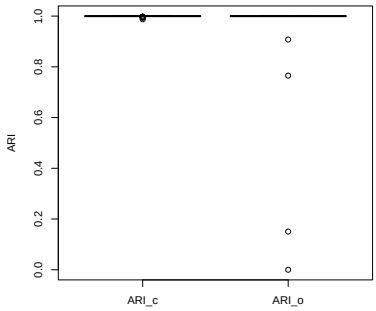

Figure 2: Performance du clustering (ARI_c) et de la détection d'anomalies (ARI_o).

3.2 Résultats

Nous fixons $K = 3$ mais notons que K peut être déterminé en considérant plusieurs valeurs et en choisissant le meilleur par BIC (Amovin et al. 2020). Les dimensions intrinsèques des clusters sont obtenues par le test de Cattell avec un seuil fixé à 0.2. Nous simulons 50 jeux de données de taille 1000. La Table 1 donne les moyennes et écarts-types des estimations des paramètres du modèle. La performance du modèle en terme de clustering et de détection d'anomalie est évaluée par l'indice de Rand (*Adjusted Rand index*, noté ARI) (Figure 2). Les résultats sont très bons, l'algorithme détectant toutes les données anormales pour 92% des simulations, sans faux positifs.

Paramètres	Groupe 1	Groupe 2	Groupe 3
a_k	150.45 (5.72)	100.22 (2.23)	52.13 (1.27)
b_k	4.91 (0.06)	2.80 (0.03)	9.59 (0.09)
d_k	5 (0)	20 (0)	10 (0)

Table 1: Moyenne (et écart type) des estimations des paramètres du modèle sur 50 simulations.

4 Conclusion

Dans ce travail, nous proposons une technique de clustering pour détecter des anomalies dans les données fonctionnelles multivariées. Les résultats obtenus à travers ce modèle sur données simulées présentent de bonnes performances. Ces résultats sont confirmés sur données réelles. Pour une mise en oeuvre opérationnelle d'un point de vue industriel de cet algorithme, nous pensons étendre ce modèle à une version *online* permettant la mise à jour des paramètres au fur et à mesure que de nouvelles données arrivent.

Bibliographie

Amovin-Assagba, M., Gannaz, I., & Jacques, J. (2021). Outlier detection in multivariate functional data through a contaminated mixture model. *arXiv preprint arXiv:2106.07222*.

Garcia-Escudero, L. A., & Gordaliza, A. (2005). A proposal for robust curve clustering. Journal of classification, 22(2), 185-201.

Hubert, M., Rousseeuw, P., & Segaert, P. (2017). Multivariate and functional classification using depth and distance. Advances in Data Analysis and Classification, 11(3), 445-466.

Punzo, A., & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, 58(6), 1506-1537.

Ramsay, J. O., & Silverman, B. W. (2005). Functional data analysis. Springer series in statistics.

Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., & Martin, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 35(3), 1101-1131.

Staerman, G., Mozharovskyi, P., Clémençon, S., & d'Alché-Buc, F. (2019, October). Functional isolation forest. In Asian Conference on Machine Learning (pp. 332-347). PMLR.