Détection d'anomalies basée sur la forme dans les données fonctionnelles multivariées

Clément Lejeune*,***, Josiane Mothe**,***
A. Soubki*,
Olivier Teste***

*Airbus Operations
prénom.nom@airbus.com,
** INSPE
*** IRIT, UMR5505 CNRS, Université de Toulouse, UT2J
prénom.nom@irit.fr

1 Introduction

Les données fonctionnelles sont définies comme des vecteurs de grande dimension contenant des mesures dépendant d'une variable continue. Par exemple, les études longitudinales mesurent un paramètre un grand nombre de fois et à différents temps de mesures pour divers individus ; ces données peuvent être considérées comme des réalisations d'une fonction univariée dépendant du temps. Ainsi, la variable aléatoire générant chaque échantillon est une fonction univariée du temps, voir les monographies Ramsay et Silverman (2006); Ferraty et Vieu (2006) pour une introduction à l'analyse des données fonctionnelles (AFD). Par extension, les données fonctionnelles multivariées sont générées par un système impliquant plusieurs fonctions, c-à-d un vecteur, dépendant du temps.

Dans le contexte des données fonctionnelles multivariées, la détection d'échantillons aberrants (anomalies), où un échantillon est une fonction donc, doit prendre en compte à la fois le comportement individuel des variables et les corrélations dynamiques entre elles ; cela rend ce problème difficile. Les corrélations entre les variables sont intéressantes à analyser car elles peuvent révéler le comportement aberrant du processus sous-jacent, comme le montre (Hubert et al., 2015).

Aggarwal et Yu (2001) définissent une valeur aberrante comme un échantillon très différent des autres, sur la base d'une mesure d'anormalité à définir. Un échantillon aberrant contient souvent des informations concernant le comportement anormal du système décrit par les données. La détection des valeurs aberrantes vise donc à déterminer une mesure appropriée permettant de différencier les valeurs aberrantes de celles qui ne le sont pas, avec un haut degré d'interprétabilité.

Les algorithmes typiques de classification comme la régression logistique et de regroupement comme K-means et mean-shift ne sont pas efficaces dans la mesure où les valeurs aberrantes sont à la fois rare et dispersées (Japkowicz et Stephen, 2002).

La détection des valeurs aberrantes dans les données fonctionnelles a principalement été étudiée dans le cas univarié (Fraiman et Muniz, 2001; Cuevas et al., 2006; López-pintado et

Romo, 2009) ; plus récemment dans le cas multivarié (Claeskens et al., 2014; Ieva et Paganoni, 2013; López-pintado et al., 2014; Hubert et al., 2015; Kuhnt et Rehage, 2016; Dai et Genton, 2019).

Les anomalies fonctionnelles multivariées sont caractérisées par des déviations dans la corrélation entre les variables et/ou dans leur corrélation avec le temps. Hubert et al. (2015) définissent deux classes générales : les aberrations isolées qui correspondent à un comportement extrême sur une petite partie du domaine et les aberrations persistantes qui sont des échantillons dans lesquels la valeur aberrante se manifeste dans une grande partie du domaine. Parmi ces dernières, les aberrations de forme présentent des caractéristiques locales aberrantes sans s'écarter des courbes régulières en tout point du domaine temporel.

L'étude de la détection des formes aberrantes est assez récente (López-pintado et al., 2014; Arribas-Gil et Romo, 2014; Kuhnt et Rehage, 2016; Dai et Genton, 2019).

Les méthodes de l'état de l'art se basent sur la profondeur statistique, qui peut être vue comme un score de centralité de chaque échantillon par rapport au reste du jeu de données fonctionnelles. Dai et Genton (2019) proposent une profondeur statistique fonctionnelle basée sur la direction que prend la fonction (vectorielle) à chaque instant t observée. Cette approche $Dir.out$, ne considère que certaines relations aberrantes, entre les variables de la fonction, dans la mesure où seulement les directions ponctuelles de chaque fonction sont prises et non sa forme globale. Kuhnt et Rehage (2016) proposent une profondeur statistique fonctionnelle qui, pour chaque variable, est basée sur l'angle, à un instant donné, entre la tangente d'un échantillon et celle d'un autre $FUNTA$. L'opération est répétée pour toutes les combinaisons d'échantillons deux-à-deux, toutes les variables et tous les instants de temps observées. Intuitivement, il s'agit de caractériser l'anormalité de chaque échantillon fonctionnel par ses variations d'angles par rapport aux autres échantillons. Cette approche ne considère pas non plus l'aberration de forme globale décrite par chaque fonction.

Les formes aberrantes persistantes sont difficiles à détecter dans une population de courbes car elles sont souvent discriminantes de façon no-linéaire et présentent une plus grande variabilité que les aberrations isolées. Pour discriminer les courbes en termes de forme, il est possible d'utiliser des outils de géométrie différentielle, c'est à dire rajouter des dérivées ou des intégrales calculées par rapport à t pour chaque variable. Ainsi, la forme de la courbe fournit des informations sur les caractéristiques aberrantes cachées des variables de la courbe et sur la relation aberrante entre elles. Cependant, l'analyse conjointe des variables devient alors complexe.

Cet article résume notre contribution, Lejeune et al. (2020), dans laquelle nous considérons ces deux aspects en agrégeant de façon géométrique les variables de chaque courbe. Le reste de l'article est organisé comme suit : en Section 2, nous résumons la représentation fonctionnelle que nous nécessitons pour ensuite calculer les fonctions d'agrégation que nous avons proposées. Nous résumons les résultats des études expérimentales menées sur des jeux de données réelles et artificielles en Section 3, puis nous concluons en Section 4.

2 Analyse de données fonctionnelles - fonctions d'agrégation

Contrairement aux méthodes actuelles de détection des anomalies fonctionnelles qui se concentrent sur la profondeur statistique fonctionnelle, dans l'étude publiée dans Lejeune et al. (2020), nous avons abordé ce problème en utilisant la géométrie différentielle de chaque échan-

tillon. Nous utilisons des fonctions d'agrégation des variables de l'échantillon fonctionnel. Ainsi, nous considérons implicitement la corrélation des variables à travers les caractérisations géométriques de la forme des courbes.

L'analyse fonctionnelle des données repose sur la représentation de vecteurs de mesures de grande dimension, dépendant du temps, par des fonctions (voir Ramsay et Silverman (2006); Ferraty et Vieu (2006)). Représenter ce type de données, comme les séries temporelles, par des fonctions permet de recouvrir la vraie nature du processus sous-jacent à la fonction qui a généré les données. Cela permet également de lisser des courbes dans le cas d'échantillons bruités. Le cadre AFD permet de traiter des courbes qui sont échantillonnées de manière irrégulière ou sur des grilles de différentes tailles, où une grille fait référence à la discrétisation d'un intervalle fermé dans lequel se trouve la variable continue dépendante (exemple : temps, longueur d'onde).

Nous utilisons à la fois les caractéristiques fonctionnelles de la forme de la courbe et les algorithmes de détection des d'anomalies comme Isolation Forest Liu et al. (2008) et One-class SVM Schölkopf et al. (2001). Ainsi, l'originalité de l'approche proposée réside dans la caractérisation de la forme des courbes initiales par le biais des fonctions de d'agrégation proposées, combinées aux algorithmes de détection d'anomalies de l'état de l'art.

Dans notre terminologie, la fonction d'agrégation désigne une fonction analytique qui permet de capturer les caractéristiques de la forme de la courbe, telles que la courbure, la longueur, ou la vitesse tangentielle, et de prendre en compte toutes les variables, une courbe étant considérée comme une trajectoire spatiale. Ces fonctions agrègent les variables, pour chaque instant observé, d'une fonction multivariée en une fonction univariée. Ces agrégats correspondent à différentes combinaisons interprétables des dérivées des variables par rapport au temps. Ce type d'approche est inspirée des techniques d'analyse des formes extraites d'images (Srivastava et Klassen, 2016).

Pour capturer le caractère aberrant des courbes à travers leur forme, nous proposons des fonctions d'agrégation inspirées de la géométrie différentielle (Srivastava et Klassen, 2016). Ces agrégations exigent que les courbes soient lisses, ce qui est rarement le cas des données brutes, souvent bruitées lorsqu'elles sont échantillonnées. Nous utilisons la représentation des données fonctionnelles pour assurer le lissage des courbes. Nous représentons chaque courbe multivariée par la courbe univariée résultant d'une agrégation puis nous utilisons les algorithmes de détection d'anomalie de l'état de l'art (Isolation Forest, One-class SVM) sur ces nouvelles courbes .

La première étape de l'AFD consiste à approximer une fonction lisse inconnue qui sous-tend l'échantillon considéré par une autre fonction d'approximation lisse au moyen de mesures discrètes et bruitées : il s'agit de l'étape d'approximation fonctionnelle. Son but est de supprimer le bruit et d'obtenir une représentation analytique de cette fonction, permettant ainsi des évaluations précises de fonctions dérivées fonctions intégrales que nous utilisons (sous réserve que la fonction d'approximation soit différentiable).

La seconde étape concerne les fonctions d'agrégation. Nous avons étudié trois fonctions d'agrégation :

— Longueur d'arc : considère la longueur cumulée de la courbe entre l'instant initial observé jusqu'à un point arbitraire.
— Vitesse : la norme Euclidienne du vecteur tangent (vecteur de dérivée première par rapport au temps de chaque variable) à la courbe. Cette fonction d'agrégation permet

d'analyser les variations instantanées de la courbe par rapport à la variable dépendante continue. Lorsqu'il s'agit du temps, elle mesure la vitesse de déplacement d'un point sur la courbe.

— Courbure : elle se rapporte à la façon dont une courbe est "courbée", ou géométriquement, le degré auquel une courbe s'écarte de la ligne tangente en un point donné.

3 Résultats et discussions

Nous avons mené une étude expérimentale sur des jeux de données réelles et synthétiques.

Les résultats ont été évalués en termes de taux de vrai positif (proportion de valeurs aberrantes correctement détectées), de taux de faux positifs (proportion de valeurs aberrantes faussement détectées) et avec l'aire sous la courbe ROC (Receiver Operating Characteristic) ainsi obtenue (Area Under the Curve AUC).

Nous avons testé l'approche proposée sur trois jeux de données. Le premier a également été utilisé par Dai et Genton (2019). Il est constitué de séries temporelles d'électrocardiogrammes (ECG) de l'activité électrique (tension) des changements cardiaques (Goldberger et al., 2000). Pour un total de 810 séries temporelles, 208 correspondent à des cas anormaux et 602 à des cas normaux. Toutes les séries temporelles sont de même taille. PenDig, le second ensemble est constitué de 10992 séries temporelles bivariées représentant des chiffres de stylo (Dua et Graff, 2017). Les chiffres sont étiquetés en fonction de leur classe, de 0 à 9. Chaque chiffre possède 8 points d'observation régulièrement échantillonnés sur les coordonnées horizontales et verticales. Nous avons sur-échantillonné cet ensemble par interpolation linéaire à 200 points sur les deux coordonnées pour se ramener à des vecteurs de mesure de grande dimension. Le troisième jeux de données sont des données simulées selon les cinq modèles proposés par Dai et Genton (2019) avec des niveaux de contamination des données variés pour simuler les valeurs aberrantes de différentes natures, d'amplitude ou de forme, persistantes ou isolées. Les détails du cadre expérimental sont donnés dans notre article original (Lejeune et al., 2020).

Nos modèles ont été comparés à deux méthodes état de l'art : la méthode $FUNTA$ proposée par Kuhnt et Rehage (2016) et la méthode $Dir.out$ proposée par Dai et Genton (2019).

Sur le jeu de données ECG, les fonctions d'agrégation Vitesse et Courbure sont les plus efficaces ; elles surpassent les références. Pour le jeu de données PenDig, notre approche est toujours plus performante que les méthodes de référence en termes de valeur de AUC ; elle permet de détecter les anomalies de forme de façon plus efficace que les méthodes de l'état de l'art. Les résultats sur les données synthétiques confirment que les méthodes de l'état de l'art sont efficaces sur des anomalies isolées mais que notre approche est meilleure sur les anomalies de forme.

Cet article, issu d'une de nos publications précédentes au journal KBS Lejeune et al. (2020) présente les résultats d'une nouvelle méthode dans laquelle les caractéristiques aberrantes sont capturées sur la base de fonctions d'agrégation issues de la géométrie différentielle.

L'étude expérimentale sur des ensembles de données réelles et synthétiques et la comparaison avec des méthodes basées sur la profondeur fonctionnelle montrent que la méthode proposée combinée aux algorithmes de détection des aberrations les plus récents peut être plus efficace. Elle est efficace quelle que soit la proportion d'aberrations.

4 Conclusion

Cet article présente les résultats de nos travaux publiés dans (Lejeune et al., 2020) dans lequel nous avons proposé une méthode pour améliorer la détection de différents types de valeurs aberrantes dans les données fonctionnelles multivariées. Notre approche se base sur la forme des courbes contrairement à l'état de l'art qui se base sur la profondeur statistique fonctionnelle considérant la forme de façon limitée.

Notre approche s'appuie sur la reconstruction lisse des courbes, par projection dans une base de fonctions fixée, puis sur l'utilisation de fonctions d'agrégation, longueur d'arc, vitesse et courbure, pour capturer les caractéristiques de forme latentes. Les valeurs aberrantes des courbes transformées sont détectées à l'aide d'algorithmes de détection de la littérature.

L'étude expérimentale sur des données réelles et des données synthétiques montre que l'approche proposée surpasse les méthodes de l'état de l'art sur les données réelles ainsi que sur les données synthétiques pour les anomalies de formes persistantes. Notre approche est par ailleurs robuste à la variation du taux de contamination.

Références

Aggarwal, C. C. et P. S. Yu (2001). Outlier Detection for High-Dimensional Data. In *SIGMOD*, Volume 30, pp. 37–46. ACM.

Arribas-Gil, A. et J. Romo (2014). Shape outlier detection and visualization for functional data : The outliergram. *Biostatistics 15*(4), 603–619.

Claeskens, G., M. Hubert, L. Slaets, et K. Vakili. (2014). MFHD : Multivariate Functional Halfspace Depth. *Journal of the American Statistical Association 109*(505), 411–423.

Cuevas, A., M. Febrero, et R. Fraiman (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis 51*, 1063–1074.

Dai, W. et M. G. Genton (2019). Directional outlyingness for multivariate functional data. *Computational Statistics and Data Analysis 131*, 50–65.

Dua, D. et C. Graff (2017). UCI machine learning repository.

Ferraty, F. et P. Vieu (2006). *Nonparametric functional data analysis : theory and practice*. Springer Science & Business Media.

Fraiman, R. et G. Muniz (2001). Trimmed means for functional data. *Test 10*(2).

Goldberger, A. L., L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, et H. E. Stanley (2000). Physiobank, physiotoolkit, and physionet : components of a new research resource for complex physiologic signals. *Circulation 101*(23), e215–e220.

Hubert, M., P. J. Rousseeuw, et P. Segaert (2015). Multivariate functional outlier detection. *Statistical Methods and Applications 24*(2), 177–202.

Ieva, F. et A. M. Paganoni (2013). Depth Measures for Multivariate Functional Data. *Communications in Statistics - Theory and Methods 42*(7), 1265–1276.

Japkowicz, N. et S. Stephen (2002). The class imbalance problem : A systematic study. *Intelligent data analysis 6*(5), 429–449.

Kuhnt, S. et A. Rehage (2016). An angle-based multivariate functional pseudo-depth for shape outlier detection. *Journal of Multivariate Analysis 146*, 325–340.

Lejeune, C., J. Mothe, A. Soubki, et O. Teste (2020). Shape-based outlier detection in multivariate functional data. *Knowledge-Based Systems 198*, 105960.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2008). Isolation Forest. In *ICDM*, pp. 413–422.

López-pintado, S. et J. Romo (2009). On the Concept of Depth for Functional Data. *Journal of the American Statistical Association 104*(486), 718–734.

López-pintado, S., Y. Sun, J. K. Lin, et M. G. Genton (2014). Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification 8*(3), 321–338.

Ramsay, J. et B. W. Silverman (2006). *Functional Data Analysis*. Wiley Online Library.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, et R. C. Williamson (2001). Estimating the support of a high-dimensional distribution. *Neural computation 13*(7), 1443–1471.

Srivastava, A. et E. P. Klassen (2016). *Functional and Shape Data Analysis*. Springer Series in Statistics.

Summary

Multivariate functional data are generated by a system involving dynamic parameters depending on continuous variables. Outlier detection must consider both the individual behavior of the parameters and the dynamic correlation between them. Recent work has focused on multivariate functional depth. These approaches fail when the outlyingness is manifested in the shape of the curve rather than in its magnitude. This paper, based on our publication in the journal KBS Lejeune et al. (2020) presents the results of a new method in which outlying features are captured based on mapping functions from differential geometry. Experimental study on real and synthetic datasets and comparison with functional depth-based methods show that the proposed method combined with the latest outlier detection algorithms can be more effective. It is effective regardless of the proportion of outliers.