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