Sujet de thèse : Analyse de données fonctionnelles multivariées hétérogènes

Alain Duhamel (alain.duhamel@univ-lille2.fr),
Cristian Preda (cristian.preda@polytech-lille.fr),
Vincent Vandewalle (vincent.vandewalle@univ-lille2.fr)

Ce sujet explore le caractère multivarié des données fonctionnelles. Il s'agit de l'étude d'une variable aléatoire fonctionnelle $X = (X_1, X_2, \ldots, X_p)$ à valeurs dans un espace produit (des espaces de fonctions). Les composantes X_i, i=1,...,p, peuvent être des variables fonctionnelles univariées (fonctions réelles) ou des variables fonctionnelles qualitatives (processus de sauts). On s'intéresse principalement à la classification non-supervisée de ce type de données. Des problématiques en classification supervisée seront aussi abordées. Plusieurs points sont à traiter lors de ce travail :

1) Réaliser un état de l'art sur l'analyse des données fonctionnelles multivariées. On trouve en littérature principalement des données fonctionnelles multivariées *scalaires*. Aspect important à traiter : visualisation de ce type de données.
2) Etat de l'art sur les données fonctionnelles qualitatives. Voir notamment les travaux de Saporta, Deville, Boumaza et les plus récentes (Preda et Vandewalle). Limites de l'utilisation de la modélisation markovienne pour ce type de données. Trajectoires des longueurs différentes – états absorbants. Visualisation.
3) Hétérogénéité des composantes X_i (scalaire/qualitatif). Quel modélisation choisir ? Le problème n'est pas tranché même dans le cas non-fonctionnel. Peut-on voir une composante qualitative comme une chaine de markov (non-caché du coup, puisque observable) qui gouverne le comportement des composantes scalaires ? Et si plusieurs composantes sont qualitatives ?
4) Développement des méthodes de classification pour données fonctionnelle multivariées : méthodes factorielles et modèles génératifs (via l'algorithme EM). Visualisation.
5) Application sur des données hospitalières, notamment sur des parcours des patients à l'hôpital. Volume de données important. Echantillonnage (travaux de H. Cardot).

Références :

R. Boumaza. *Contribution a l'étude descriptive d'une fonction aléatoire qualitative. PhD thesis, Université Paul Sabatier, Toulouse, France, 1980.*

Jean-Claude Deville, *Analyse harmonique du calendrier de constitution des familles en France. Disparités sociales et évolution de 1920 à 1960, Population, Volume 32, Numéro 1, pp. 17-63, 1977.*

Vincent Vandewalle, Cristina Cozma et Cristian Preda, *Clustering categorical functional data Application to medical discharge letters, 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, Londres, 2015.*