Richesse et complexité des données fonctionnelles

Frédéric Ferraty[1], Philippe Vieu[2]

[1] Auteur correspondant : Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse Cedex, France, ferraty@cict.fr

[2] Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse Cedex, France

Résumé Les progrés récents en matière de stockage et de traitement des données se traduisent de plus en plus fréquemment dans de nombreux domaines scientifiques par la présence de données de type fonctionnel (courbes, images, ...). Les défis proposés aux statisticiens pour appréhender ce type de données ont abouti depuis quelques années à la construction de nombreuses méthodes statistiques. Il se trouve que la complexité de ce type de données amène une richesse d'information qu'une méthode statistique (aussi sophistiquée soit elle) arrive difficilement à capter, tandis que des techniques de boosting capables d'utiliser les complémentarités de différentes méthodes se révèlent souvent plus performantes. L'objectif de ce travail est d'illustrer ce point de vue au travers d'un problème couramment rencontré en pratique : celui de la prévision d'une variable réponse réelle à partir d'une variable explicative fonctionnelle. Un rapide tour d'horizon des méthodes habituellement utilisées sera effectué, et leur complémentarité sera mise en évidence au travers d'un jeu de données issu d'un problème de chimie quantitative.

Keywords : Analyse de données fonctionnelles, Boosting, Méthodes de sélection, Modèles fonctionnels, Régression, Spectrométrie, Statistique non-paramétrique.

1 Introduction

La plupart des domaines scientifiques font face à des situations où les données recueillies sont de nature continue (courbes, images, ...). On pourrait citer par exemple, sans chercher à être exhaustif, la biologie, la climatologie, l'économétrie, la chimie quantitative, ... Bien évidemment, ces données continues ne sont en réalité observées que sur une grille formée par un ensemble fini de points de discrétisation, de telle sorte que l'on peut penser à les traiter au moyen des outils usuels (paramétriques ou non) de la statistique multidimensionnelle. Cette approche naïve des choses peut s'avérer intéressante dans des situations où les points de discrétisation sont peu nombreux et relativement distants les uns des autres. Depuis quelques années les moyens technologiques en matière de recueil (et de stockage) de données se sont considérablement accrus, amenant des grilles d'observations de données fonctionnelles de plus en plus fines et rendant inadéquates les méthodes statistiques multivariées usuelles, non seulement du fait des problèmes de grande dimension liés au nombre important de variables mais aussi du fait de la corrélation importante existant entre deux observations proches d'un même phénomène continu. Cette situation paradoxale conduirait à considérer que, loin d'être un avantage pour la connaissance des phénomènes, l'abondance de données aboutirait à une détérioration des résultats statistiques. Ce paradoxe était tellement fort et présent chez les statisticiens dans les années