Proposition de stage de M2 en statistique mathématique

Présentation

Titre du stage : Estimation dans le modèle linaire fonctionnel avec réponse fonctionnelle.

Contacts : Gaëlle Chagny (CR, LMRS UMR CNRS 6085 , Univ. Rouen) et Angelina Roche (MCF, CEREMADE UMR CNRS 7534, Univ. Paris Dauphine),

Courriels : gaelle.chagny@univ-rouen.fr et roche@ceremade.dauphine.fr

Lieu du stage : Laboratoire de Mathématiques Raphaël Salem, Univ. Rouen.

Durée : 4 mois.

Contexte : Le stage se déroulera dans l'équipe partenaire de Rouen du projet ANR SMILES (Statistical Modeling and Inference for unsupervised Learning at largE-Scale), porté par Faïcel Chamrouki (Professeur, LMNO, Univ. Caen). Ce projet est principalement dévolu à la modélisation et l'inférence statistique pour des données complexes, de grande échelle ("Big data"), via des problèmes de régression et de classification non-supervisée. Le sujet du stage s'insère dans la partie analyse de données fonctionnelles du projet.

Financement : ANR.

Description scientifique

Mots Clés. Analyse de données fonctionnelles. Modèle linéaire fonctionnel. Statistique non-paramétrique.

Cadre. Ce stage a trait à l'analyse statistique de données fonctionnelles : il s'agit de l'étude d'observations qui ne sont pas, comme généralement en statistique, des réalisations de variables aléatoires réelles ou vectorielles (vecteurs aléatoires), mais des fonctions aléatoires (courbes, images, etc...). Ce sont des données de dimension infinie, c'est-à-dire rentrant dans le champ de la "très grande dimension". Celles-ci apparaissent de plus en plus fréquemment dans de nombreux domaines scientifiques, grâce aux progrès récents en matière de stockage et traitement. La biologie, la climatologie, l'économétrie ou encore la chimie sont par exemple susceptibles de produire des données considérées comme des courbes aléatoires. Leur traitement requiert des méthodes spécifiques, différentes (ou tout au moins spécifiquement adaptées) de celles de l'analyse statistique multivariée classique.

Les recherches en statistique pour données fonctionnelles se sont multipliées ces dernières décennies : on pourra par exemple consulter l'une des nombreuses monographies sur le sujet, comme celles de Ramsay et Silverman (2002, 2005); Ferraty et Vieu (2006); Ferraty et Romain (2011).

Au niveau mathématique, les connaissances requises sont à l'interface de l'analyse fonctionnelle, des probabilités et des statistiques.

Problème étudié. Un problème très classique en statistique est celui de l'étude du lien entre une variable d'intérêt Y et une covariable ou variable explicative X. Le modèle sous-jacent, dit de régression, peut généralement s'écrire

$$Y = r(X) + \varepsilon,$$

où r modélise le lien (à estimer) entre X et Y et ε, un bruit, aléatoire également.

Le stage se concentre sur le cas où les deux variables aléatoires X et Y sont fonctionnelles : elles sont considérées comme étant à valeurs dans un espace de Hilbert séparable $(\mathbb{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$, typiquement l'espace $L^2(I)$ des fonctions de carré intégrable sur un intervalle I de $\mathbb{R}$ ou un espace de Sobolev plus général. On se place donc non plus dans le cadre fini-dimensionnel de la statistique classique mais

dans un cadre infini-dimensionnel. On supposera que le lien entre X et Y est linéaire, au sens où il existe un opérateur linéaire $S : \mathbb{H} \to \mathbb{H}$ tel que $r(X) = SX$. Typiquement, dans le cas où $\mathbb{H} = L^2(I)$, S sera un opérateur intégral, au sens où

$$SX(t) = \int_{[0,1]} \mathcal{S}(s,t)X(s)ds,\ t \in [0,1]$$

pour un certain noyau intégrable $\mathcal{S}$. L'objectif est d'estimer S (ou $\mathcal{S}$), à partir d'un échantillon $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ ($n \in \mathbb{N}\backslash\{0\}$) distribué selon la loi de (X, Y). Le champs couvert est donc à la fois celui de la statistique pour données fonctionnelles, et celui de la statistique non-paramétrique, puisque l'opérateur S entre X et Y, bien que linéaire, vit lui aussi dans un espace de dimension infinie. Un tel problème généralise

- d'une part l'étude du modèle linéaire fonctionnel, où seule la variable X est supposée fonctionnelle (la variable réponse Y étant scalaire), introduit par Ramsay et Dalzell (1991) et largement étudié d'un point de vue théorique et appliqué depuis le travail précurseur de Cardot *et al.* (1999) (voir par exemple Cardot *et al.* 2003; Crambes *et al.* 2009; Cardot et Johannes 2010; Brunel *et al.* 2016),
- d'autre part l'étude d'autres problèmes de régression fonctionnelle où seule la variable d'intérêt Y est fonctionnelle (le design X étant multivarié), voir par exemple la revue de Chiou *et al.* (2004).

Motivé pourtant par des applications pratiques (étude de la consommation d'électicité par exemple, Antoch *et al.* 2010; Benatia *et al.* 2017), peu de résultats théoriques semblent exister sur le modèle linéaire fonctionnel avec réponse fonctionnelle, à l'exception des travaux de Yao *et al.* (2005) et Crambes et Mas (2013), où des études asymptotiques d'estimateurs fondés sur l'ACP fonctionnelle de X sont proposés.

Objectifs du stage. Dans un premier temps, l'objectif du stage sera de faire un point sur les méthodes existant dans la littérature sur le modèle considéré. Dans un second temps, il sera possible d'étudier d'un point de vue non-asymptotique un estimateur de la fonction S (par exemple celui proposé par Crambes et Mas 2013 ou un estimateur à noyau) : l'écriture d'une décomposition de type biais-variance pour un risque à définir pourra permettre de proposer une méthode de sélection de la dimension (ou de la fenêtre) inspirée des travaux de Birgé et Massart (1998) ou Goldenshluger et Lepski (2011).

La bibliographie ci-dessous n'est pas exhaustive, elle se concentre juste sur quelques éléments en liaison directe avec le programme de recherche.

Références

J. Antoch, L. Prchal, M. R. De Rosa et P. Sarda : Electricity consumption prediction with functional linear regression using spline estimators. *J. Appl. Stat.*, 37(12):2027–2041, 2010.

D. Benatia, M. Carrasco et J.-P. Florens : Functional linear regression with functional response. *J. Econometrics*, 201(2):269–291, 2017.

L. Birgé et P. Massart : Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

E. Brunel, A. Mas et A. Roche : Non-asymptotic adaptive prediction in functional linear models. *J. Multivariate Anal.*, 143:208–232, 2016. ISSN 0047-259X.

H. Cardot, F. Ferraty et P. Sarda : Functional linear model. *Statist. Probab. Lett.*, 45(1):11–22, 1999.

H. Cardot, F. Ferraty et P. Sarda : Spline estimators for the functional linear model. *Statist. Sinica*, 13(3):571–591, 2003.

H. Cardot et J. Johannes : Thresholding projection estimators in functional linear models. *J. Multivariate Anal.*, 101(2):395–408, 2010.

J.-M. Chiou, H.-G. Müller et J.-L. Wang : Functional response models. *Statist. Sinica*, 14(3):675–693, 2004.

C. Crambes, A. Kneip et P. Sarda : Smoothing splines estimators for functional linear regression. *Ann. Statist.*, 37(1):35–72, 2009.

C. Crambes et A. Mas : Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, 19(5B):2627–2651, 2013.

F. Ferraty et Y. Romain : *The Oxford Handbook of Functional Data Analysis*. Oxford Handbooks in Mathematics. OUP Oxford, 2011.

F. Ferraty et P. Vieu : *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

A. Goldenshluger et O. Lepski : Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

J. O. Ramsay et C. J. Dalzell : Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B*, 53(3):539–572, 1991. With discussion and a reply by the authors.

J. O. Ramsay et B. W. Silverman : *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. Methods and case studies.

J. O. Ramsay et B. W. Silverman : *Functional data analysis*. Springer Series in Statistics. Springer, New York, second édn, 2005.

F. Yao, H.-G. Müller et J.-L. Wang : Functional linear regression analysis for longitudinal data. *Ann. Statist.*, 33(6):2873–2903, 2005.