Sujet de stage de M2 en statistique
Étude de l'impact de la discrétisation pour l'estimation du modèle linéaire fonctionnel

Présentation

Encadrement du stage.

- Gaëlle Chagny (gaelle.chagny@univ-rouen.fr), Université de Rouen Normandie,
- Vincent Rivoirard (Vincent.Rivoirard@dauphine.fr), Université Paris Dauphine,
- Angelina Roche (angelina.roche@u-paris.fr), Université Paris Cité (anciennement Paris-Descartes).

Contexte. Projet ANR FUNMathStat : Mathematical Statistics for Functional Data Analysis.

Ce projet est principalement dévolu aux problématiques spécifiques soulevées par l'analyse des données fonctionnelles du point de vue de la statistique mathématique, à l'interface entre la statistique des processus et l'analyse fonctionnelle.

Site du projet : `https://sites.google.com/view/funmathstatanrproj/accueil`.

Lieu du stage. CEREMADE, Université Paris-Dauphine/LMRS Université de Rouen-Normandie.

Une gratification est envisagée pour le stage et un financement de thèse est prévu dans la continuité du stage (date de début envisagée pour la thèse septembre ou octobre 2025).

Résumé

Contexte. Ce sujet de 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 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 dans ce domaines 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 Romain (2011); Horváth et Kokoszka (2012).

La plupart des contributions théoriques sur le sujet supposent que les données fonctionnelles sont observées complètement : le postulat est typiquement celui de l'observation de courbes aléatoires sur tout un intervalle de $\mathbb{R}$. Or, en pratique, nous ne disposons en général que de données discrétisées, c'est-à-dire observées sur des grilles finies, et potentiellement bruitées. Les grilles d'observation peuvent être fixes ou aléatoires, régulières ou non, communes à tous les individus ou propre à chacun. L'étape préalable de lissage, permettant de passer de ces observations discrètes aux courbes aléatoires, n'est que rarement prise en compte dans l'étude théorique des procédures de statistique inférentielle mises en œuvre ensuite.

L'objectif de ce stage est d'étudier l'impact de la discrétisation des données sur la qualité de l'estimation dans un des modèles les plus étudiés en analyse de données fonctionnelles : le modèle linéaire fonctionnel à sortie scalaire.

Cadre statistique. Nous nous intéressons à la modélisation statistique du lien entre une variable explicative fonctionnelle X, considérée comme élément d'un espace de Hilbert séparable $\mathbb{H}$ muni d'un produit scalaire $\langle \cdot, \cdot \rangle$ et une variable réponse réelle Y. Le modèle le plus classique, introduit sous la forme ci-dessous par Cardot *et al.* (1999), propose une relation linéaire entre X et Y,

$$Y = \beta_0 + \langle \beta, X \rangle + \varepsilon, \tag{1}$$

où $\beta_0 \in \mathbb{R}$ est l'intercept, $\beta \in \mathbb{H}$ la fonction de pente (*slope function*), et ε un bruit centré indépendant de X. L'estimation du couple (β_0, β) à partir d'un échantillon de données $(X_i, Y_i)_{i=1,\dots,n}$ distribuées comme (X, Y) est un problème inverse largement étudié dans la littérature. Les méthodes proposées sont fondées sur une régularisation du problème, par projection des données dans une base de fonctions de référence (base orthonormée quelconque, base trigonométrique, base de splines, base de l'ACP, base PLS ...) et/ou par minimisation d'un critère de type moindres carrés pénalisés (pénalisation de type ridge par exemple, ou méthode de type splines de lissage). Les contrôles de risques d'estimation associés (bornes de risques non-asymptotiques de type décomposition biais-variance, vitesses de convergence, bornes inférieures du risque minimax...) sont obtenus sous l'hypothèse d'une observation complète des covariables $(X_i)_{i=1,\dots,n}$, au sens où, si par exemple $\mathbb{H} = \mathbb{L}^2(\mathcal{T})$ pour un intervalle $\mathcal{T}$ de $\mathbb{R}$, on suppose observer $\{X_i(t),\, t \in \mathcal{T}\},\ i = 1, \dots, n$: voir par exemple Cai et Hall (2006); Cardot et Johannes (2010); Comte et Johannes (2012); Brunel *et al.* (2016), et les références citées dans ces articles.

En pratique, les covariables sont plutôt observées sous la forme discrétisée $\{W_i(t_j),\, (i,j) \in \{1,\dots,n\} \times \{1,\dots,p\}\}$ où $(t_j)_{j=1,\dots,p}$ est un ensemble de points de discrétisation de l'intervalle $\mathcal{T}$, et $W_i(t_j)$ une évaluation bruitée de $X_i(t_j)$:

$$W_i(t_j) = X_i(t_j) + \eta_{i,j}, \tag{2}$$

les variables $(\eta_{i,j})$ étant i.i.d. centrées, de variance finie, indépendantes des $(X_i)_i$. L'objectif du stage est de comprendre comment cette discrétisation, traitée en pratique par une étape préalable de lissage, influe en théorie sur les vitesses d'estimation du couple (β_0, β). Comment, à partir de l'échantillon observé $\{Y_i, W_i(t_j),\, (i,j) \in \{1,\dots,n\} \times \{1,\dots,p\}\}$ peut-on adapter les procédures d'estimation usuelles par projection dans une base orthonormée ? Quelles vitesses de convergence, dépendant cette fois des deux indices p et n atteignent ces procédures ? Quelles vitesses minimax est-on en droit d'attendre pour ce problème de régression ?

Bien que très importante en pratique, ce type de problématique apparaît assez peu dans la littérature, et principalement pour d'autres problèmes d'estimation : Cai et Yuan (2011) ont calculé des vitesses minimax, pour l'estimation de la fonction moyenne d'un échantillon de données fonctionnelles, à la fois dans le cas d'une grille fixe, et d'une grille aléatoire. Le cas de l'estimation des éléments propres de l'opérateur de covariance (fonctions propres et valeurs propres empiriques) a été étudié par Hall *et al.* (2006) (cas d'une grille aléatoire), et plus récemment par Belhakem *et al.* (2024) (estimation de la première fonction propre). Une revue de l'ensemble des résultats récents est proposée par Roche (2022). Concernant le modèle linéaire fonctionnel, proposé pour le stage, Cardot *et al.* (2007) introduisent la régression par splines de lissage, et obtiennent des vitesses de convergence tenant compte de la discrétisation des données, mais qui ne sont pas minimax. Crambes (2007) propose une étape de lissage des covariables par méthodes à noyau, puis étudie la consistance de l'estimateur de β obtenu par régression en composantes principales. Mais la question de l'obtention de bornes supérieures et inférieures d'estimation dans le modèle (1) à partir des covariables bruitées (2) reste à notre connaissance un problème ouvert.

Le premier axe de recherche consistera à définir et étudier un estimateur par projection dans la base des fonctions propres de l'opérateur de covariance de X, base que l'on pourra supposer connue dans un premier temps (comme par exemple Brunel et Roche 2015), à partir des observations

$\{Y_i, W_i(t_j),\ (i,j) \in \{1,\dots,n\} \times \{1,\dots,p\}\}$. Les résultats pourront ensuite être étendus à une régression en composantes principales plus générale, et complétés par une borne inférieure d'estimation.

De nombreuses extensions de ce travail seront ensuite possibles. Du point de vue de la modélisation tout d'abord, il est envisageable de considérer un modèle linéaire dont la sortie est également fonctionnelle : la prise en compte de deux grilles d'observations non nécessairement identiques (l'une pour les covariables, l'autre pour la réponse) sera considérée. Du point de vue théorique, le cas d'une grille d'observation aléatoire pourrait également être intéressant.

La bibliographie ci-dessous donne quelques éléments indicatifs.

Références

R. Belhakem, F. Picard, V. Rivoirard et A. Roche : Minimax estimation of functional principal components from noisy discretized functional data. *Scandinavian Journal of Statistics*, 2024.

É. Brunel, A. Mas et A. Roche : Non-asymptotic adaptive prediction in functional linear models. *Journal of Multivariate Analysis*, 143:208–232, 2016.

E. Brunel et A. Roche : Penalized contrast estimation in functional linear models with circular data. *Statistics*, 49(6):1298–1321, 2015.

T. T. Cai et P. Hall : Prediction in functional linear regression. *The Annals of Statistics*, p. 2159–2179, 2006.

T. T. Cai et M. Yuan : Optimal estimation of the mean function based on discretely sampled functional data : Phase transition. *The Annals of Statistics*, 39(5):2330 – 2355, 2011.

H. Cardot, C. Crambes, A. Kneip et P. Sarda : Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational statistics & data analysis*, 51(10):4832–4848, 2007.

H. Cardot, F. Ferraty et P. Sarda : Functional linear model. *Statistics & Probability Letters*, 45 (1):11–22, 1999.

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

F. Comte et J. Johannes : Adaptive functional linear regression. *The Annals of Statistics*, 40 (6):2765–2797, 2012. ISSN 0090-5364,2168-8966.

C. Crambes : Régression fonctionnelle sur composantes principales pour variable explicative bruitée. *Comptes Rendus. Mathématique*, 345(9):519–522, 2007.

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

P. Hall, H.-G. Müller et J.-L. Wang : Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493 – 1517, 2006.

L. Horváth et P. Kokoszka : *Inference for Functional Data with Applications*. Springer, New York, 2012.

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.

A. Roche : New perspectives in smoothing : minimax estimation of the mean and principal components of discretized functional data. *The Graduate Journal of Mathematics*, 7(2):95–107, 2022.