Publications du Groupe de Travail STPAH :
Statistique Fonctionnelle et Opératorielle

STAPH 2010-02 *Résumés des exposés aux sessions de Statistique Fonctionnelle. Journées de Statistique, Marseille. 25-28 mai 2010*

Alain Boudou, Frédéric Ferraty, Yves Romain, Pascal Sarda, Philippe Vieu et Sylvie Viguier-Pla

Institut de Mathématiques, Université Paul Sabatier, Toulouse, France

TABLE DES MATIÈRES

Présentation

Alain BOUDOU, Frédéric FERRATY, Yves ROMAIN, Pascal SARDA, Philippe VIEU et Sylvie VIGUIER-PLA

Groupe de travail STAPH
Institut de Mathématiques de Toulouse
118 Route de Narbonne 31062 Toulouse Cedex

Le groupe de travail STAPH a initié depuis juin 2002 des journées annuelles de Statistique Fonctionnelle et Opératorielle. Ces journées offrent la possiblité à des jeunes chercheurs ou des chercheurs confirmés de laboratoires français ou étrangers de présenter leurs travaux et d'échanger avec d'autres chercheurs. Un des objectifs est de mettre en lumière la diversité d'approches et de travaux balayant le spectre de la théorie et des applications. Les premières éditions de ces journées se sont déroulées à Toulouse et à partir de 2006 dans d'autres villes universitaires : Grenoble (2006), Lille (2007) et Dijon (2009). La tenue de ces journées a été suspendue en 2008 pour faire place au premier Workshop international de Statistique Fonctionnelle et Opératorielle à Toulouse, IWFOS'08. La seconde édition de IWFOS, qui est dans sa phase préparatoire, se tiendra à Santander en juin 2011. Nous nous réjouissons du succès rencontré par ces manifestations et du réseau de recherche croissant qu'elles ont contribué à créer, générant ainsi de nombreuses collaborations, nouveaux développements ...

Cette année les rencontres STAPH se sont déroulées dans le cadre inhabituel des journées de statistique de Marseille qui se sont tenues du 25 au 28 mai 2011. L'initiative en revient à David Nerini. Nous le remercions chaleureusement ici pour avoir rendu possible l'organisation de deux sessions de Statistique Fonctionnelle au sein des journées de Statistique.

On trouvera dans le document présent l'ensemble des résumés des exposés lors de ces sessions. Comme on pourra s'en rendre compte, les travaux exposés reflètent une pluralité d'approches théoriques et/ou appliquées, illustrant le dynamisme de la recherche en Statistique Fonctionnelle dans ses différents aspects. Nous remercions l'ensemble des orateurs pour leurs contributions.

L'ensemble des activités du groupe de travail STAPH peut être consulté sur la page :

http ://www.math.univ-toulouse.fr/staph/

Méthode de lissage bayésienne tempérée pour estimer les paramètres d'un modèle d'équation différentielle

David CAMPBELL* et Russell STEELE

* Adresse pour correspondance :
Department of Statistics and Actuarial Science,
Simon Fraser University,
Surrey BC, Canada, V3T 0A3

e-mail : dac5@sfu.ca

Résumé

L'utilisation répandue des modèles d'équations différentielles ordinaires (EDO) a depuis longtemps été sous-représentée dans la littérature statistique. Les méthodes les plus communes pour estimer les paramètres des modèles d'EDO sont les moindres carrés nonlinéaires [1] et une méthode basée sur les MCMC [2]. Ces méthodes dépendent d'une vraisemblance basée sur la solution numérique de l'EDO. Le défi relevé par ces méthodes est que les espaces de paramètres sont difficiles à naviguer, aggravé par la grande variété de formes fonctionnelles qu'un modèle d'EDO peut produire avec des petits changements de valeurs des paramètres. Bien que certains progrès récents ont été accomplis dans la littérature fréquentiste grâce à l'utilisation du lissage (par exemple [3],[4]), les méthodes bayésiennes n'ont pas suivi. Ce travail décrit la méthode de lissage bayésienne tempérée (LBT). Cette méthode utilise une expansion de bases pour approximer la solution d'EDO dans la vraisemblance, où la forme de l'expansion est guidée par le modèle d'EDO. Cette approximation de l'EDO lisse la surface de vraisemblance, réduisant ainsi les restrictions de mouvement des paramètres. La méthode de LBT utilise une suite de densités postérieures basée sur des approximations lisses à la solution d'EDO. Le niveau de l'approximation est déterminé par la valeur du paramètre de lissage qui contrôle le niveau de rugosité dans la surface de vraisemblance. Dans un algorithme semblable au tempérant parallèle, des chaînes MCMC parallèles sont utilisées pour échantillonner la suite de densités postérieures, tout en permettant aux paramètres d'EDO de permuter entre les chaînes. La combinaison de méthodes bayésienne et de méthodes pour les données fonctionelle améliore la convergence, tout en permettant l'inférence sur des vraisemblances identiques

aux modèles utilisés par les moindres carrés non-linéaires et une méthode basée sur les MCMC traditionnels.

Mots-clés : Données Fonctionelles, Méthodes bayésiennes, Modèles semi et non paramétriques, Lissage.

Abstract

The widespread use of ordinary differential equation (ODE) models has long been under-represented in the statistical literature. The most common methods for estimating parameters from ODE models are nonlinear least squares [1] and an MCMC based method [2]. Both of these methods depend on a likelihood involving the numerical solution to the ODE. The challenge faced by these methods is parameter spaces that are difficult to navigate, exacerbated by the wide variety of behaviours that a single ODE model can produce with respect to small changes in parameter values. While frequentist literature has seen some recent improvements in methodology thanks to the incorporation of smoothing methods (for example [3],[4]), Bayesian methods have not yet followed. This work describes a new Bayesian method, Smooth Functional Tempering, using a basis expansion to approximate the ODE solution in the likelihood, where the shape of the basis expansion, or data smooth, is guided by the ODE model. This approximation to the ODE, smooths out the likelihood surface, reducing restrictions on parameter movement. Smooth Functional Tempering, uses a sequence of posterior densities with smooth approximations to the ODE solution. The level of the approximation is determined by the value of the smoothing parameter, which also determines the level of smoothness in the likelihood surface. In an algorithm similar to parallel tempering, parallel MCMC chains are run to sample from the sequence of posterior densities, while allowing ODE parameters to swap between chains. The incorporation of smoothing methods improves convergence while ultimately enabling inference on the same likelihoods as are used in traditional MCMC methods. This method is introduced and tested against a variety of alternative Bayesian models, in terms of posterior variance and rate of convergence.

KeyWords : Functional Data, Bayesian Methods, Semi-parameteric modelling, Smoothing

Références

[1] Bates, D. M., and Watts, D. B. (1988). *Nonlinear Regression Analysis and Its Appli-*

cations, Wiley books, New York.

[2] Gelman, A., Bois, F. Y., and Jiang, J. (1996). Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions, *Journal of the American Statistical Association*, **91**, 1400-1412.

[3] Liang, H., and Wu, H. (2008). Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models, *Journal of the American Statistical Association*, **103**, 1570-1583.

[4] Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007), Parameter Estimation for Differential Equations :A Generalized Smoothing Approach (with Discussion), *Journal of the Royal Statistical Society Series B*, **69**, 741-796.

Semiparametric Models with Functional Responses in a Survey Sampling Setting : Model assisted Estimation of Electricity Consumption Curve

Hervé CARDOT, Alain DESSERTAINE et Etienne JOSSERAND*

* Adresse pour correspondance :
Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France

e-mail : Etienne.Josserand@u-bourgogne.fr

Résumé

Ce travail adopte une approche de type sondage quand le but est d'estimer une courbe moyenne d'une grande base de données de données fonctionnelles. Lorsque les capacités de stockage sont limitées, grâce aux techniques de sondage, une petite partie des observations est une alternative intéressante par rapport aux techniques de compression. Nous proposons ici de prendre en considération une information auxiliaire réelle ou multivariée obtenu à moindre coût sur la population toute entière, avec une approche semiparamétrique de type modèle assisté, dans le but d'améliorer les estimateurs d'Horvitz-Thompson de la courbe moyenne. D'abord, nous estimerons les composantes principales afin de réduire la dimension des signaux, et ensuite nous utiliserons des modèles semiparamétriques pour estimer les courbes qui n'ont pas été observées. Cette technique se montre vraiment efficace sur une base de données réelle de 18902 courbes de consommation électrique mesurée toutes les demi heures pendant deux semaines.

Abstract

This work adopts a survey sampling point of view when one has to estimate the mean curve of large databases of functional data. When storage capacities are limited selecting

with survey techniques a small fraction of the observations is an interesting alternative to signal compression techniques. We propose here to take account of real or multivariate auxiliary information available at a low cost for the whole population, with semiparametric model assisted approaches, in order to improve the accuracy of Horvitz-Thompson estimators of the mean curve. We first estimate the functional principal components with a design based point of view in order to reduce the dimension of the signals and then propose semiparametric models to get estimations of the curves that are not observed. This technique is shown to be really effective on a real dataset of 18902 electricity meters measuring every half an hour electricity consumption over two weeks.

Key words : Design-based estimation, Functional Principal Components, Horvitz-Thompson estimator

1. Introduction

With the development of distributed sensors one can have access of potentially huge databases of signals evolving along fine time scales. Collecting in an exhaustive way such data would require very high investments both for transmission of the signals through networks as well as for storage. As noted in Chiky and Hébrail (2009) survey sampling procedures on the sensors, which allow a trade off between limited storage capacities and accuracy of the data, can be relevant approaches compared to signal compression in order to get accurate approximations to simple estimates such as mean or total trajectories.

Our study is motivated, in such a context of distributed data streams, by the estimation of the temporal evolution of electricity consumption curves. The French operator EDF has planned to install in a few years more than 30 millions electricity meters, in each firm and household, that will be able to send individual electricity consumptions at very fine time scales. Collecting, saving and analysing all this information which can be seen as functional would be very expensive and survey sampling strategies are interesting to get accurate estimations at reasonable costs (Dessertaine, 2006). It is well known that consumption profiles strongly depend on covariates such as past consumptions, meteorological characteristics (temperature, nebulosity, *etc*) or geographical information (altitude, latitude and longitude). Taking this information into account at an individual level (*i.e* for each electricity meter) is not trivial. One way to achieve this consists in reducing first the high dimension of the data by performing a functional principal components analysis in a survey sampling framework with a design based approach (Cardot *et al.*, 2010). It is then possible to build models, parametric or nonparametric, on the principal component scores in order to incorporate the auxiliary variables effects and correct our estimator with model assisted approaches (Särndal *et al.*, 1992). Note that this strategy based on modeling the principal components instead of the original signal has already been proposed, with a frequentist point of view, by Chiou *et al.* (2003) with singel index models and

Müller and Yao (2008) with additive models.

We present in section 2 the Horvitz-Thomposon estimator of the mean consumption profile as well as the functional principal components analysis. We develop, in section 3, model assisted approaches based on statistical modeling of the principal components scores and derive an approximated variance that can be useful to build global confidence bands. Finally, we illustrate, in section 4, this methodology which allows to improve significantly more basic approaches on a population of 18000 electricity consumption curves measured every half an hour over one week.

2. Functional data in a finite population

Let us consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ with size N, and suppose we can observe, for each element k of the population U, a deterministic curve $Y_k = (Y_k(t))_{t\in[0,1]}$ that is supposed to belong to $L^2[0,1]$, the space of square integrable functions defined on the closed interval $[0,1]$ equipped with its usual inner product $\langle\cdot,\cdot\rangle$ and norm denoted by $\|\cdot\|$. Let us define the mean population curve $\mu \in L^2[0,1]$ by

$$\mu(t) \;=\; \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,1]. \tag{1}$$

Consider now a sample s, *i.e.* a subset $s \subset U$, with known size n, chosen randomly according to a known probability distribution p defined on all the subsets of U. We suppose that all the individuals in the population can be selected, with probabilities that may be unequal, $\pi_k = \Pr(k \in s) > 0$ for all $k \in U$ and $\pi_{kl} = \Pr(k \;\&\; l \in s) > 0$ for all $k, l \in U$, $k \neq l$.

The Horvitz-Thompson estimator of the mean curve, which is unbiased, is given by

$$\widehat{\mu}(t) \;=\; \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{Y_k(t)}{\pi_k}\mathbb{1}_{k\in s}, \quad t \in [0,1]. \tag{2}$$

As in Cardot *et al.* (2010) we would like to describe now the individual variations around the mean function in a functional space whose dimension is as small as possible according to a quadratic criterion. Let us consider a set of q orthonormal functions of $L^2[0,1]$, $\phi_1, \ldots, \phi_q$, minimize, according to $\phi_1, \ldots, \phi_q$, the remainder $R(q)$ of the projection of the Y_k's onto the space generated by these q functions

$$R(q) = \frac{1}{N}\sum_{k\in U}\|R_{qk}\|^2$$

with

$$R_{qk}(t) = Y_k(t) - \mu(t) - \sum_{j=1}^{q}\langle Y_k - \mu, \phi_j\rangle\phi_j(t), \quad t \in [0,1].$$

Introducing now the population covariance function $\gamma(s,t)$,

$$\gamma(s,t) = \frac{1}{N}\sum_{k\in U}\left(Y_k(t)-\mu(t)\right)\left(Y_k(s)-\mu(s)\right), \quad (s,t)\in[0,1]\times[0,1],$$

Cardot *et al.* (2010) have shown that $R(q)$ attains its minimum when $\phi_1,\ldots,\phi_q$ are the eigenfunctions of the covariance operator Γ associated to the largest eigenvalues, $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_q\geq 0$,

$$\Gamma\phi_j(t) = \int_0^1 \gamma(s,t)\phi_j(s)ds = \lambda_j\phi_j(t), \quad t\in[0,1], j\geq 1.$$

When observing individuals from a sample s, a simple estimator of the covariance function

$$\widehat{\gamma}(s,t) \;=\; \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\left(Y_k(t)-\widehat{\mu}(t)\right)\left(Y_k(s)-\widehat{\mu}(s)\right)\;\; (s,t)\in[0,1]\times[0,1], \qquad (3)$$

allows to derive directly estimators of the eigenvalues $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_q$ and the corresponding eigenfuctions $\widehat{\phi}_1,\ldots,\widehat{\phi}_q$.

3. Semiparametric estimation with auxiliary information

Suppose now we have access to m auxiliary variables $X_1,\ldots,X_m$ that are supposed to linked to the individual curves Y_k and we are able to observe these variables, at a low cost, for every individual k in the population. Taking this additional information into account would certainly be helpful to improve the accuracy of the basic estimator $\widehat{\mu}$. Going back to the decomposition of the individual trajectories Y_k on the eigenfunctions,

$$Y_k(t) = \mu(t) + \sum_{j=1}^{q}\langle Y_k-\mu,\phi_j\rangle\phi_j(t) + R_{qk}(t), \quad t\in[0,1],$$

and borrowing ideas from Chiou *et al.* (2003) and Müller and Yao (2008), an interesting approach consists in modeling the population principal components scores $\langle Y_k-\mu,\phi_j\rangle$ with respect to auxiliary variables at each level j of the decomposition on the eigenfunctions,

$$\langle Y_k-\mu,\phi_j\rangle \approx f_j(x_{k1},\ldots,x_{km})$$

where the regression function f_j can be parametric or not and $(x_{k1},\ldots,x_{km})$ is the vector of observations of the m auxiliary variables for individual k.

It is possible to estimate the principal component scores $\widehat{C}_{kj} = \langle Y_k-\widehat{\mu},\widehat{\phi}_j\rangle$, for $j=1,\ldots,q$ and all $k\in s$ and then build a design based least squares estimator for the functions f_j

$$\widehat{f_j} \;=\; \arg\min_{g_j}\sum_{k\in s}\frac{1}{\pi_k}\left(\widehat{C}_{kj}-g_j(x_{k1},\ldots,x_{km})\right)^2, \qquad (4)$$

in order to construct the following model-assisted estimator $\widehat{\mu}_X$ of μ :

$$\widehat{\mu}_x(t) \;=\; \widehat{\mu}(t) - \frac{1}{N}\left(\sum_{k\in s}\frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k\in U}\widehat{Y}_k(t)\right) \tag{5}$$

where the predicted curves $\widehat{Y}_k$ are estimated for all the individuals of the population U thanks to the m auxiliary variables,

$$\widehat{Y}_k(t) = \widehat{\mu}(t) + \sum_{j=1}^{q}\widehat{f}_j(x_{k1},\ldots,x_{km})\;\widehat{v}_j(t),\;\; t\in[0,1].$$

4. Application : estimation of electricity consumption curves

We have a population of $N = 18902$ electricity meters that are able to send electricity consumptions every half an hour over a period of two weeks, so that we have $d = 336$ time points. We are interested in estimating the mean consumption curve over the second week and we suppose that we know the mean consumption, $\bar{Y}_k = \frac{1}{336}\sum_{j=1}^{336} Y_k(t_j)$, for each meter k of the population over the first week. This mean consumption will play the role of auxiliary information. Note that meteorological variables are not available in this preliminary study.

We first perform a simple random sampling without replacement (SRSWR) with fixed size of $n = 2000$ electricity meters over the second week order to get $\widehat{\mu}$ and perform the functional principal components analysis (FPCA).

To evaluate the accuracy of estimator (5) we made 500 replications of the following scheme

- Draw a sample of size $n = 2000$ in population U with SRSWR and estimate $\widehat{\mu}, \widehat{\phi}_1$ and $\widehat{C}_{k1}$, for $k \in s$, over the second week.
- Estimate a linear relationship between X_k and $\widehat{C}_{k1}$, for $k \in s$ where $X_k = \frac{1}{336}\sum_{j=1}^{336} Y_k(t_j)$ is the mean consumption over the first week, $\widehat{C}_{k1} \approx \widehat{\beta}_0 + \widehat{\beta}_1 X_k$.
- Estimate $\widehat{\mu}_X$ taking the auxiliary information into account with equation (5).

The following loss criterion $\int |\mu(t) - \widehat{\mu}(t)|dt$ has been considered to evaluate the accuracy of the estimators $\widehat{\mu}$ and $\widehat{\mu}_X$.

We will present the detail results during the presentation and we will show that model assisted estimators allow a significant improvement compared to the basic SRSWR approach.

Acknowledgment. Etienne Josserand thanks the Conseil Régional de Bourgogne for its financial support (FABER PhD grant).

References

[1] CARDOT, H., CHAOUCH, M., GOGA, C. and C. LABRUÈRE (2010). Properties of Design-Based Functional Principal Components Analysis, *J. Statist. Planning and Inference.*, **140**, 75-91.

[2] CARDOT, H., JOSSERAND, E. (2009). Horvitz-Thompson Estimators for Functional Data : Asymptotic Confidence Bands and Optimal Allocation for Stratified Sampling. `http ://arxiv.org/abs/0912.3891`.

[3] CHIKY, R., HEBRAIL, G. (2009). Spatio-temporal sampling of distributed data streams. *J. of Computing Science and Engineering*, to appear.

[4] CHIOU, J-M., MÜLLER, H.G. and WANG, J.L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J.Roy. Statist. Soc., Ser. B*, **65**, 405-423.

[5] DESSERTAINE, A. (2006). Sondage et séries temporelles : une application pour la prévision de la consommation électrique. *38èmes Journées de Statistique*, Clamart, Juin 2006.

[6] MÜLLER, H-G., YAO, F. (2008). Functional Additive Model. *J. Am. Statist. Ass.* **103**, 1534-1544.

[7] SÄRNDAL, C.E., SWENSSON, B. and J. WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

[8] SKINNER, C.J, HOLMES, D.J, SMITH, T.M.F (1986). The Effect of Sample Design on Principal Components Analysis. *J. Am. Statist. Ass.* **81**, 789-798.

Prédiction en régression linéaire fonctionnelle avec variable d'intérêt fonctionnelle

Christophe CRAMBES* et André MAS

* Adresse pour correspondance :
Université Montpellier 2,
Place Eugène Bataillon,
34095 Montpellier Cedex, France
e-mail : ccrambes@math.univ-montp2.fr

Résumé. Ce travail concerne l'étude de la prédiction dans le modèle linéaire fonctionnel lorsque la variable d'intérêt est elle aussi fonctionnelle. Nous introduisons un prédicteur basé sur les décompositions de Karhunen-Loève des courbes X (variable explicative) et Y (variable d'intérêt). Les résultats obtenus permettent de fournir un développement asymptotique de la moyenne quadratique de l'erreur de prédiction. Nous donnons également un résultat d'optimalité pour ces vitesses dans un sens minimax, ainsi qu'un théorème de la limite centrale du prédicteur.

Abstract. This work concerns the prediction problem in the functional linear model with functional output. We introduce a predictor based on Karhunen-Loève decompositions of the curves X (covariate) and Y (output). Our results give an asymptotic development of the mean square prediction error. We also give an optimality result for these rates of convergence in a minimax sense, as well as a central limit theorem for the predictor.

Mots clés. Modèle linéaire fonctionnel, variable d'intérêt fonctionnelle, décomposition de Karhunen-Loève, erreur de prédiction, vitesses optimales, théorème de la limite centrale.

1. Introduction

Les modèles de régression, permettant d'expliquer comment une variable d'intérêt Y est reliée à une variable explicative X, sont parmi les plus utilisés en statistique. Nous nous plaçons dans ce cadre de travail, en supposant que les variables X et Y sont à valeurs dans l'espace $L^2(I)$ des fonctions de carré intégrable sur un intervalle I, qui sera considéré comme $[0,1]$ pour simplifier. Ce type de variables aléatoires dites fonctionnelles permet de prendre en compte de nombreuses situations pratiques où les observations sont par nature des courbes (fonctions du temps par exemple). Ces données étant très présentes dans de nombreuses applications, les travaux concernant l'étude des données fonctionnelles se multiplient actuellement à très grande vitesse. Les ouvrages de référence actuels en la matière sont les monographies de Ramsay et Silverman (2002, 2005), qui donnent une vue d'ensemble sur ce champ de recherche, tandis que la monographie de Ferraty et Vieu (2006) recense les principaux résultats obtenus dans un contexte non-paramétrique sur les données fonctionnelles.

On considère dans la suite un modèle qui s'écrit sous la forme

$$Y(t) = \int_0^1 \mathcal{S}(s,t)\, X(s)\, ds + \varepsilon(t), \quad \mathbb{E}(\varepsilon|X) = 0, \tag{6}$$

où $\mathcal{S}$ est un noyau intégrable. Ce modèle, encore peu étudié, a fait l'objet de quelques travaux, parmi lesquels Chiou, Müller et Wang (2004), Yao, Muller et Wang (2005) qui proposent une estimation de $\mathcal{S}$ basée sur une analyse en composantes principales des courbes X et Y. Une des premières études est due à Cuevas, Febrero et Fraiman (2002). Récemment, Antoch *et al.* (2008) ont étudié un estimateur spline de $\mathcal{S}$ tandis que Aguilera, Ocaña and Valderrama (2008) en ont proposé un estimateur à base d'ondelettes. Le modèle (6) peut s'écrire sous la forme $Y(t) = S(X)(t) + \varepsilon(t)$ où l'opérateur S est défini par $Sf(t) = \int_0^1 \mathcal{S}(s,t)\, f(s) ds$ pour toute fonction f de L^2.

Dans la suite, on considère un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$ d'observations indépendantes et de même loi que (X,Y) sur lequel on se base pour construire notre prédicteur.

2. Construction du predicteur

On introduit les notations suivantes. Le produit scalaire usuel de L^2 est noté $\langle .,. \rangle$ et défini par $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$ pour toutes fonctions f et g de L^2. Le produit tensoriel entre deux fonctions f et g de L^2 est défini par $f \otimes g = \langle g, . \rangle f$ et associe à toute fonction h de L^2 la fonction $\langle g, h \rangle f$. Partant du modèle (6), il vient

$$\mathbb{E}[Y \otimes X] = \mathbb{E}[S(X) \otimes X] + \mathbb{E}[\varepsilon \otimes X].$$

En notant

$$\Delta = \mathbb{E}\left[Y \otimes X\right], \quad \Gamma = \mathbb{E}\left[X \otimes X\right],$$

on en déduit $\Delta = S\Gamma$. En introduisant les versions empiriques des opérateurs Δ et Γ par

$$\Delta_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \otimes X_i, \quad \Gamma_n = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i,$$

un estimateur naturel de S est donné par $\widehat{S}_n$ vérifiant $\Delta_n = \widehat{S}_n\Gamma_n$. Le problème est que l'opérateur Γ_n ne peut pas être directement inversé. Une solution classique consiste à considérer un inverse régularisé. Pour cela, on note $\left(\widehat{\lambda}_j, \widehat{e}_j\right)$ les éléments propres de Γ_n (les valeurs propres étant rangées par ordre décroissant). De façon analogue, (λ_j, e_j) désignent les éléments propres de Γ. L'opérateur Γ_n s'écrit alors $\Gamma_n = \sum_j \widehat{\lambda}_j \left(\widehat{e}_j \otimes \widehat{e}_j\right)$ et son inverse régularisé est donné par

$$\Gamma_n^{\dagger} = \sum_{j=1}^{k} \widehat{\lambda}_j^{-1} \left(\widehat{e}_j \otimes \widehat{e}_j\right), \tag{7}$$

où $k = k_n$ est le nombre de composantes principales choisies. De cette décomposition se déduit une expression de l'estimateur de $\mathcal{S}$ par

$$\widehat{\mathcal{S}}_n\left(s,t\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{\int X_i \widehat{e}_j}{\widehat{\lambda}_j} Y_i\left(t\right)\widehat{e}_j\left(s\right).$$

L'estimateur $\widehat{S}_n$ de S est défini par $\widehat{S}_n = \Delta_n\Gamma_n^{\dagger}$ et le prédicteur associé est donné par $\widehat{Y}_{n+1} = \widehat{S}_n(X_{n+1}) = \Delta_n\Gamma_n^{\dagger}(X_{n+1})$ pour une nouvelle observation X_{n+1}.

3. Résultats asymptotiques

3.1. Hypothèses

Les hypothèses permettant d'établir nos résultats sont les suivantes.

$(H.1)$ On suppose que S est un opérateur de Hilbert-Schmidt : pour toute base $(e_j)_{j\in\mathbb{N}}$ de H, on a

$$\sum_{j,\ell} \left\langle S\left(e_\ell\right), e_j\right\rangle^2 < +\infty.$$

(H.2) Considérons la décomposition de Karhunen-Loève de X qui s'écrit

$$X = \sum_{j=1}^{+\infty} \sqrt{\lambda_j} \xi_j e_j \quad p.s.,$$

où les ξ_j sont des variables aléatoires centrées réduites et non corrélées. On suppose que, pour $j, \ell \in \mathbb{N}$, il existe une constante b telle que

$$\mathbb{E}\left(|\xi_j|^\ell\right) \leq \frac{\ell!}{2} b^{\ell-2} \cdot \mathbb{E}\left(|\xi_j|^2\right).$$

(H.3) Soit λ la fonction définie par $\lambda(j) = \lambda_j$ pour tout entier j (les λ_j étant les valeurs propres de l'opérateur Γ). On interpole cette fonction de façon continue entre j et $j+1$ telle que

$$x \to \lambda(x) \text{ est convexe.}$$

L'hypothèse (H.1) équivaut à supposer que le noyau $\mathcal{S}$ est doublement intégrable. Remarquons qu'en dehors de cette hypothèse, on ne suppose rien d'autre sur S, en particulier aucune hypothèse de régularité n'est requise. L'hypothèse (H.2) a comme conséquence de faire une hypothèse de moment (d'ordre 4) sur X. Cette hypothèse est par exemple vérifiée lorsque X est un processus gaussien ou encore un processus borné p.s. L'hypothèse (H.3) est un hypothèse de décroissance sur les valeurs propres de Γ. Elle est vérifiée pour une large classe d'opérateurs, dont les valeurs propres sont à décroissance arithmétique, exponentielle ..., y compris pour des processus X très irréguliers.

3.2. Erreur de prédiction en moyenne quadratique

On note $\Gamma_\varepsilon = \mathbb{E}(\varepsilon \otimes \varepsilon)$ l'opérateur de covariance du bruit et $\sigma_\varepsilon^2 = \mathrm{tr}\Gamma_\varepsilon$. On a alors, sous les hypothèses précédentes

$$\mathbb{E}\left\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\right\|^2 = \sigma_\varepsilon^2 \frac{k}{n} + \sum_{j=k+1}^{+\infty} \lambda_j \|S(e_j)\|^2 + A_n + B_n, \tag{8}$$

où $A_n \leq C_A \frac{k^2 \lambda_k}{n} \|S\|_{\mathcal{L}_2}$ et $B_n \leq C_B \frac{k^2 (\log k)}{n^2}$, les constantes C_A et C_B ne dépendant pas de k, n ou S.

3.3. Optimalité

Notre estimateur est construit de façon très proche de celui de Yao, Muller et Wang (2005). Notre apport concerne un résultat, énoncé ci-dessous, donnant des vitesses optimales de convergence de l'estimateur. Pour toute fonction $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ de classe C^1 et décroissante telle que $\sum_{j=1}^{+\infty} \varphi(j) = 1$, on note $\mathcal{L}_2(\varphi, L)$ la classe des opérateurs linéaires de H dans H définie par

$$\mathcal{L}_2(\varphi, L) = \left\{ T \in \mathcal{L}_2, \|T\|_{\mathcal{L}_2} \leq L : \|T(e_j)\| \leq L\sqrt{\varphi(j)} \right\}.$$

Si on note de plus $L = \left\|S\Gamma^{1/2}\right\|_{\mathcal{L}_2}$, $\varphi(j) = \lambda_j \|S(e_j)\|^2 / L^2$ et k_n^* la partie entière de la solution en x de l'équation

$$\frac{1}{x}\int_x^{+\infty} \varphi(x)\, dx = \frac{1}{n}\frac{\sigma_\varepsilon^2}{L^2},$$

on a alors (comme conséquence de la sous-section précédente) :

$$\lim\sup_{n\to+\infty} \frac{n}{k_n^*} \sup_{S\Gamma^{1/2}\in\mathcal{L}_2(L,\varphi)} \mathbb{E}\left\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\right\|^2 = 2\sigma_\varepsilon^2.$$

Le résultat d'optimalité peut à présent être énoncé :

$$\inf_{\widehat{S}_n} \sup_{S\in\mathcal{L}_2(\varphi,C)} \mathbb{E}\left\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\right\|^2 \asymp \frac{k_n^*}{n}.$$

3.4. Convergence faible

Notre principal résultat est donné ci-dessous. Sous les hypothèses précédentes et sous la condition que $(k\log k)^2 / n \to 0$, alors

$$\sqrt{\frac{n}{k}}\left[\widehat{S}_n(X_{n+1}) - S\Pi_k(X_{n+1})\right] \xrightarrow{w} \mathcal{G}_\varepsilon$$

où $\mathcal{G}_\varepsilon$ est une variable aléatoire gaussienne à valeurs dans H, centrée et d'opérateur de covariance Γ_ε. Sous certaines conditions, ce résultat peut alors s'écrire

$$\sqrt{\frac{n}{k}}\left[\widehat{S}_n(X_{n+1}) - S(X_{n+1})\right] \xrightarrow{w} \mathcal{G}_\varepsilon.$$

Une des conséquences de ce résultat est que, pour un choix de $H = W_0^{2,1}([0,1]) = \{f \in L^2([0,1]) : f(0) = 0, f' \in L^2([0,1])\}$, on obtient, pour un t_0 fixé dans $[0,1]$

$$\mathbb{P}\left(Y_{n+1}^{*}\left(t_0\right) \in \left[\widehat{Y}_{n+1}\left(t_0\right) \pm \sqrt{\frac{k}{n}}\sigma_{t_0} q_{1-\alpha/2}\right]\right) = 1-\alpha,$$

avec $\sigma_{t_0}^2 = \Gamma_{\varepsilon}\left(t_0, t_0\right)$.

Bibliographie

[1] Aguilera, A., Ocaña, F. and Valderrama, M. (2008). Estimation of functional regression models for functional responses by wavelet approximation. *International Workshop on Functional and Operatorial Statistics 2008 Proceedings, Functional and operatorial statistics*, Dabo-Niang and Ferraty (Eds.), Physica-Verlag, Springer.

[2] Antoch, J., Prchal, L., De Rosa, M. and Sarda, P. (2008). Functional linear regression with functional response : application to prediction of electricity consumption. *International Workshop on Functional and Operatorial Statistics 2008 Proceedings, Functional and operatorial statistics*, Dabo-Niang and Ferraty (Eds.), Physica-Verlag, Springer.

[3] Chiou, J-M., Müller, H-G. and Wang J-L. (2004). Functional response models. *Statistica Sinica*, **14**, 659-677.

[4] Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression : the case of a fixed design and a functional response. *Canadian Journal of Statistics*, **30**, 285-300.

[5] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis : methods, theory, applications and implementations*. Springer-Verlag, London.

[6] Ramsay, J.O. and Silverman, B.W. (2002). *Applied functional data analysis*. Springer-Verlag.

[7] Ramsay, J.O. and Silverman, B.W. (2005). *Functional data analysis* (Second Ed.). Springer, New York.

[8] Yao F., Müller H-G. and Wang, J-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.*, **33**, 2873-2903.

Utilisation de tests de structure en régression sur variable fonctionnelle.

Laurent DELSOL*, Frédéric FERRATY et Philippe VIEU

* Adresse pour correspondance :
Université d'Orléans,
MAPMO, Fédération Denis Poisson. Route de Chartres, B.P. 6759 - 45067 Orléans cedex 2 FRANCE
e-mail : laurent.delsol@univ-orleans.fr

Abstract : This work focuses on recent advances on the way general structural testing procedures can be constructed in regression on functional variable. Our test statistic is constructed from a specific estimator adapted to the specific model to be checked and uses recent advances concerning kernel smoothing methods for functional data. A general theoretical result states the asymptotic normality of our statistic under the null hypothesis and diverges under the local alternatives. This result opens interesting prospects about tests for no-effect, for linearity, or for reduction dimension of the covariate. Bootstrap methods are then proposed to compute the threshold value of our test. Finally, we present some applications to spectrometric datasets and discuss interesting prospects for the future.

Mots-clés : test de structure, régression, variable fonctionnelle, rééchantillonnage, non-effet, linéaire, multivarié, spectrométrie.

1. Introduction. Certains phénomènes évoluent au cours du temps ou des conditions du milieu dans lequel l'expérience est réalisée. Il n'est donc pas rare d'être amené à prendre en compte des observations discrétisées de leur évolution (pouvant être modélisée par une courbe) afin d'étudier de manière plus pertinente un problème concret. Les récents progrès technologiques permettent fréquemment de disposer de données discrétisées sur des grilles assez fines qui reflètent de manière appropriée la nature fonctionnelle de ces phénomènes. De nombreuses méthodes ont été proposées afin de sélectionner parmi l'ensemble de ces observations discrétisées un petit nombre de points permettant de répondre aussi bien que

possible au problème posé. Cependant, il est souvent intéressant d'attacher également de l'importance à la dynamique de ce type de phénomènes ainsi qu'à leur structure particulière. Une manière adaptée d'y parvenir consiste à modéliser les données dont nous disposons comme la discrétisation d'une variable fonctionnelle (c'est à dire de dimension infinie). C'est une manière de généraliser l'approche multivariée qui permet d'obtenir une représentation plus synthétique des données prenant en compte la régularité et la nature intrinsèque du phénomène dont proviennent nos observations.

La branche de la statistique consacrée à l'étude de données fonctionnelles est actuellement en plein essor en raison des perspectives pratiques et théoriques qu'elle propose. De nombreux modèles et méthodes de statistique multivariée ont été généralisées afin de s'adapter à ce nouveau type de modélisation. On pourra notamment consulter les ouvrages de références de Ramsay et Silverman (1997, 2002, 2005), Bosq (2000), Ferraty et Vieu (2006), ainsi que Ferraty et Romain (2010) . Nous nous intéressons plus particulièrement dans ce travail à l'étude de problèmes de régression sur variable fonctionnelle :

$$Y = r(\mathcal{X}) + \epsilon,$$

où Y est une varable aléatoire réelle, $\mathcal{X}$ une variable aléatoire à valeurs dans un espace semi-métrique $(\mathcal{E}, d)$ et $\mathbb{E}[\epsilon|\mathcal{X}] = 0$.
De nombreux auteurs ont déjà considéré l'estimation de l'opérateur de régression r au travers de variantes de ce modèle correspondant à différentes hypothèses sur la structure de l'opérateur r. On peut notamment évoquer le modèle linéaire fonctionnel introduit par Ramsay et Dalzell (1991) :

$$Y = \alpha_0 + < \alpha, \mathcal{X} >_{\mathbb{L}^2([0;1])} + \epsilon,\ (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{L}^2([0;1]).$$

Ce modèle a été abondamment étudié au cours des dernières années comme en témoignent notamment les travaux de Cardot *et al.* (1999,2000,2007), Ramsay et Silverman (1997, 2005), Preda et Saporta (2005), Hall et Cai (2006), Crambes *et al.* (2009) ainsi que Ferraty et Romain (2010, Chapitre 2).
Divers autres modèles basés sur une certaine structure de r ont été considérés comme on peut notamment le voir dans les travaux de Sood *et al.* (2009) concernant un modèle additif multivarié basé sur les premiers coefficients d'une A.C.P. fonctionnelle, Ait Saidi *et al.* (2008) à propos du modèle à indice simple fonctionnel, ou Aneiros-Perez et Vieu (2009) pour le modèle partiellement linéaire fonctionnel. Cela illustre la grande diversité des modélisations que l'on peut proposer, d'autant plus qu'il est vraisemblable que de nouveaux exemples de modèles "structurels" soient considérés dans les années à venir (modèles additifs fonctionnels, partiellement fonctionnel, ...).
D'autre part, Ferraty et Vieu (2000) ont considéré un modèle non-pararamétrique fonctionnel dans lequel aucune hypothèse n'est faite sur la structure de r, mais simplement sur sa régularité (de type Hölder). De nombreuses références sont données à ce propos

dans les travaux de Ferraty *et al.* (2002), Masry (2005), Ferraty et Vieu (2006), Delsol (2007,2009) ainsi que Ferraty et Romain (2010, Chapitres 1, 4, et 5).

2. Tests de structure.

2.1. Généralités.

Comme nous venons de le voir, la littérature concernant les méthodes d'estimation en régression sur variable fonctionnelle est assez conséquente. L'objectif de cet exposé est sensiblement différent puisque l'on ne désire pas estimer l'opérateur r mais construire des outils statistiques permettant de tester si il a une certaine structure (e.g. constant, linéaire, multivarié, ...). La littérature consacrée à ce type de problèmes se limite, autant que nous le sachions, aux travaux de Cardot *et al.* (2003,2004) dans le cas particulier du modèle linéaire, Gadiaga et Ignaccolo (2005) qui proposent des tests de non effet basés sur des méthodes de projections ainsi que Chiou et Müller (2007) qui introduisent une approche heuristique pour construire un test d'adéquation. Il semble donc qu'il n'existe pas de méthode générale permettant de tester la validité des différents modèles évoqués au paragraphe précédant. Notons dans ce qui suit $\mathcal{R}$ une famille d'opérateurs de carré intégrables et w une fonction de poids. Au travers de cet exposé nous souhaitons présenter une approche générale permettant de tester l'hypothèse nulle

$$\mathcal{H}_0 : \{\exists r_0 \in \mathcal{R}, P(r(\mathcal{X}) = r_0(\mathcal{X})) = 1\}$$

contre des alternatives locales de la forme

$$\mathcal{H}_{1,n} : \{ \inf_{r_0 \in \mathcal{R}} \|r - r_0\|_{\mathbb{L}^2(wdP_{\mathcal{X}})} \geq \eta_n\}.$$

Notre statistique de test est construite, de manière similaire à l'approche utilisée par Härdle et Mammen (1993), à partir d'un estimateur $\hat{r}$ spécifique au modèle que l'on veut tester (donc à la famille $\mathcal{R}$) et de méthodes d'estimation à noyau (noté K) :

$$T_n = \int (\sum_{i=1}^{n} (Y_i - \hat{r}(\mathcal{X}_i)) K \left(\frac{d(\mathcal{X}_i, x)}{h_n} \right))^2 w(x) dP_{\mathcal{X}}(x).$$

Pour des raisons techniques, on fait l'hypothèse que l'estimateur $\hat{r}$ est construit sur un échantillon D_1 indépendant de $D = (\mathcal{X}, Y_i)_{1 \leq i \leq n}$ Un résultat donné par Delsol *et al.* (2010) montre la normalité asymptotique de T_n sous l'hypothèse nulle et sa divergence sous l'alternative sous des hypothèses générales. Ce résultat permet d'envisager l'utilisation de ce type de statistique de test dans un gand nombre de situations pour lesquelles les hypothèses peuvent être vérifiées comme par exemple :

– test d'un modèle a priori : $\mathcal{R} = \{r_0\}$, $\hat{r} = r_0$.

- test de non effet : $\mathcal{R} = \{r : \exists C \in \mathbb{R},\, r \equiv C\}$, $\hat{r} = \overline{Y}_n$.
- test de modèle multivarié : $\mathcal{R} = \{r : r = g \circ V,\, V : \mathcal{E} \to \mathbb{R}^p$ connu, $g : \mathbb{R}^p \to \mathbb{R}\}$, $\hat{r}$ estimateur multivarié à noyau construit à partir de $(Y_i, V(\mathcal{X}_i))_{1 \leq i \leq n}$.
- test de linéarité : $\mathcal{R} = \{r : r = \alpha_0 + < \alpha, . >, (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{L}^2[0;1]\}$, $\hat{r}$ estimateur fonctionnel spline (voir Crambes *et al.* 2009).
- test de modèle à indice simple fonctionnel : $\mathcal{R} = \{r : r = g(< \alpha, . >), \alpha \in \mathcal{E},\, g : \mathbb{R} \to \mathbb{R}\}$, $\hat{r}$ estimateur proposé par Ait Saidi *et al.* (2008).

D'autres situations peuvent également être considérées dès lors que l'on est en mesure de fournir un estimateur $\hat{r}$ ayant de bonnes propriétés.

2.2. Utilisation concrète.

La mise en oeuvre de la procédure de test décrite plus haut nécessite de calculer la valeur seuil du test. On pourrait penser l'estimer à partir de la loi asymptotique. Cependant, les termes dominants du bias et de la variance sont difficiles à estimer, c'est pourquoi on préfère utiliser des méthodes de rééchantillonnage. L'idée est de générer, à partir de l'échantillon original, B échantillons pour lesquels l'hypothèse nulle est approximativement vérifiée. Ensuite, on calcule sur chacun de ces échantillons la valeurs de la statistique de test et on prend comme valeur seuil la quantile empirique d'ordre $1 - \alpha$ des valeurs obtenues.

Nous proposons la procédure de rééchantillonnage suivante dans laquelle les étapes 1-4 sont réalisées séparément sur les échantillons $D : (\mathcal{X}_i, Y_i)_{1 \leq i \leq n}$ et $D_1 : (\mathcal{X}_i, Y_i)_{n+1 \leq i \leq N}$. Dans les lignes suivantes $\hat{r}_K$ représente l'estimateur à noyau fonctionnel de l'opérateur de régression calculé à partir de l'échantillon considéré (D or D1).

Procédure de rééchantillonnage :
Pré-traitement :

1. $\hat{\epsilon_i} = Y_i - \hat{r}_K(X_i)$
2. $\tilde{\epsilon}_i = \hat{\epsilon_i} - \bar{\hat{\epsilon}}$

Répéter B fois les étapes 3-5 :

3. Générer les résidus (3 méthodes différentes NB, SNB ou WB)
   - NB $(\epsilon_i^b)_{1 \leq i \leq n}$ tirés avec remise parmi $(\tilde{\epsilon}_i)_{1 \leq i \leq n}$
   - SNB $(\epsilon_i^b)_{1 \leq i \leq n}$ générés à partir d'une version lissée $\tilde{F}_n$ de la fonction de répartition empirique de $(\tilde{\epsilon}_i)_{1 \leq i \leq n}$ $(\epsilon_i^b = \tilde{F}_n^{-1}(U_i),\, U_i \sim \mathcal{U}(0,1))$
   - WB$(\epsilon_i^b) = \tilde{\epsilon}_i V_i$ où $V_i \sim P_W$ vérifie les conditions suivantes : $E[V_i] = 0$, $E[V_i^2] = 1$ et $E[V_i^3] = 1$.

4. Générer des réponses "correspondant" à $\mathcal{H}_0$

$$Y_i^b = \hat{r}\left(X_i\right) + \epsilon_i^b$$

5. Calculer la statistique de test T_n^b à partir de l'échantillon généré $(\mathcal{X}_i, Y_i^b)_{1\leq i\leq N}$

Calculer la valeur empirique du seuil

6. Pour un test de niveau α, prendre comme valeur le quantile empirique d'ordre $1-\alpha$ de la famille $(T_n^b)_{1\leq b\leq B}$.

On considèrere notamment trois exemples de lois P_W données par Mammen (1993). Les différentes méthodes utilisées pour générer les résidus ont des propriétés différentes. Au vu des simulations il semble intéressant d'utiliser des méthodes de type "bootstrap sauvage" (WB) qui produisent des tests plus puissants et sont par nature plus robustes à l'hétéroscédasticité des résidus.

Enfin, l'intégrale par rapport à $P_{\mathcal{X}}$ qui apparait dans la définition de T_n est approchée par une moyenne empirique sur un troisième échantillon indépendant de D_1 et D_2.

2.3. Application en spectrométrie.

Les courbes spectrométriques constituent un exemple intéressant de données de nature fonctionnelle. Elles correspondent à la mesure de l'absorption d'une lumière émise en direction d'un produit en fonction de sa longeur d'onde. Les courbes spectrométriques peuvent notamment être utilisées pour connaître le contenu d'un produit sans avoir besoin de réaliser une analyse chimique (voir par exemple Borggaard et Thodberg, 1992). Il est courant, en chimie quantitative, de faire une transformation des courbes originales (correspondant en quelque sorte à des dérivations). L'approche que nous venons de présenter peut être appliquée dans ce contexte pour apporter des éléments de réponse à des questions portant sur

- la validité d'un modèle proposé par des spécialistes.
- l'existence d'un lien entre une des dérivées de la courbe spectrométrique et la quantité que l'on cherche à prédire.
- la nature du lien reliant les dérivées de la courbe spectrométrique et le contenu chimique du produit
- la validité d'un modèle ne prenant en compte que certaines portions ou points de la courbe spectrométrique (ou de ses dérivées) dont on suppose qu'ils résument l'information apportée par la courbe spectrométrique.

Nous illustrerons brièvement la manière dont ces questions peuvent être addréssées en étudiant des jeux de données concrets.

3. Discussion.

L'approche générale que nous venons de présenter constitue une première méthode pour construire des tests de structure de nature assez variée en régression sur variable fonctionnelle (se rapporter à Delsol (2008) pour une discussion plus complète). L'utilisation de ces tests sur des données spectrométriques nous fournit des informations pertinentes sur la structure du lien entre la courbe spectrométrique et le contenu chimique du produit. De tels outils peuvent également s'avérer intéressants lorsque l'on cherche à extraire les informations pertinentes de la courbe explicative, ce qui permet souvent d'améliorer la qualité d'estimations. Il semble toutefois intéressant d'essayer d'améliorer notre approche et de proposer d'autres statistiques de test. Il serait notamment important de proposer une alternative qui ne nécessite pas de découper notre échantillon original en trois sous-échantillons, ce qui peut se révéler génant en pratique. Toutefois, il est à noter que notres approche offre un large spectre d'applications possibles. Elle pourrait être utilisée de manière intéressante dans un algorithme permettant de sélectionner les portions ou les points informatifs de la variable explicative fonctionnelle. Elle pourrait aussi se révéler intéressante dans le cadre du choix de la semi-métrique car elle peut permettre de tester la régularité de r par rapport à une semi-métrique d_1 contre la régularité de r par rapport à une semi-métrique d_2 vérifiant $d_1 \leq d_2$. Nous discuterons les éventuelles améliorations qui peuvent être effectuées pour améliorer notre approche et conclurons sur les perspectives à venir.

Bibliographie

[1] Ait-Saïdi, A., Ferraty, F., Kassa, R. et Vieu, P. (2008) Cross-validated estimations in the single functional index model. soumis

[2] Aneiros-Perez, G. and Vieu, P. (2008) Time series prediction : a semi-functional partial linear model. *Journal of Multivariate Analysis*, Accepté.

[3] Borggaard, C. et Thodberg, H.H. Optimal minimal neural interpretation of spectra, *Analytical chemistry*, 64, (5), 545–551.

[4] Bosq, D. (2000) *Linear Processes in Function Spaces : Theory and Applications*, Lecture Notes in Statistics, 149, Springer-Verlag, New York.

[5] Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003) Testing Hypothesys in the Functional Linear Model, *Scandinavian Journal of Statistics*, 30, 241–255.

[6] Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional Linear Model *Statist. and Prob. Letters*, 45, 11–22.

[7] Cardot, H., Ferraty, F. et Sarda, P. (2000) Etude asymptotique d'un estimateur spline hybride pour le modèle linéaire fonctionnel. (French) [Asymptotic study of a hybrid spline estimator for the functional linear model] *C. R. Acad. Sci. Paris*, 330, (6), 501–504.

[8] Cardot, H., Goia, A. et Sarda, P. (2004) Testing for no effect in functional linear regression models, some computational approaches. *Comm. Statist. SimulationComput.*

33, (1), 179–199.

[9] Cardot, H., Crambes, C., Kneip, A. and Sarda, P (2007) Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis, special issue on functional data analysis*, 51, (10), 4832–4848.

[10] Chiou, J.M. and Müller H.-G. (2007) Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, 51, (10), 4849–4863.

[11] Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics*, 37, 35–72.

[12] Delsol, L. (2007) Régression non-paramétrique fonctionnelle : Expressions asymptotiques des moments. *Annales de l'I.S.U.P.*, LI, (3), 43–67.

[13] Delsol, L. (2008) Régression sur variable fonctionnelle : Estimation, Tests de structure et Applications. *Thèse de doctorat de l'Université de Toulouse*.

[14] Delsol, L. (2009) Advances on asymptotic normality in nonparametric functional Time Series Analysis. *Statistics*, 43, (1), 13–33.

[15] Delsol, L., Ferraty, F., and Vieu, P. (2010) Structural test in regression on functional variables. soumis

[16] Ferraty F., Goia A. and Vieu P. (2002b) Functional nonparametric model for time series : a fractal approach for dimension reduction. *Test*, 11, (2), 317–344.

[17] Ferraty, F. et Romain, Y. (2010) *Oxford Handbook on Statistics and FDA* To appear.

[18] Ferraty, F. and Vieu, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés, *Compte Rendus de l'Académie des Sciences*, Paris, 330, 403–406.

[19] Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis : Theory and Practice* , Springer-Verlag, New York.

[20] Gadiaga, D. and Ignaccolo, R.(2005) Test of no-effect hypothesis by nonparametric regression. *Afr. Stat.*, 1, (1), 67–76.

[21] Hall, P. and Cai, T.T. (2006) Prediction in functional linear regression. (English summary) *Ann. Statist.*, 34, (5), 2159–2179.

[22] Härdle, W. and Mammen, E. (1993) Comparing Nonparametric Versus Parametric Regression Fits *Annals of Statistics*, 21, (4), 1926–1947.

[23] Masry, E. (2005) Nonparametric regression estimation for dependent functional data : asymptotic normality, *Stochastic Process. Appl.*, 115, (1), 155-177.

[24] Mammen, E. (1993) Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.*, 21, (1), 255–285.

[25] Preda, C. et Saporta, G. (2005) PLS regression on a stochastic process. *Comput. Statist. Data Anal.*, 48, (1), 149–158.

[26] Ramsay, J. and Dalzell, C. (1991) Some tools for functional data analysis, *J.R. Statist. Soc. B.*, 53, 539–572.

[27] Ramsay, J. and Silverman, B. (1997) *Functional Data Analysis*, Springer-Verlag, New York.

[28] Ramsay, J. and Silverman, B. (2002) *Applied functional data analysis : Methods and case studies*, Spinger-Verlag, New York.

[29] Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis (Second Edition)*, Spinger-Verlag, New York.

[30] Sood, A., James, G. and Tellis, G. (2009) Functional Regression : A New Model for Predicting Market Penetration of New Products *Marketing Science*, 28, 36–51.

A Functional Regression Approach for Prediction in a District-Heating System

Aldo GOIA

* Adresse pour correspondance :
Dipartimento di Scienze Economiche e Metodi Quantitativi,
Università del Piemonte Orientale A. Avogadro
Via Perrone 18, 28100 Novara, Italy
e-mail : aldo.goia@eco.unipmn.it

Résumé

Nous considérons le problème de la prédiction à court terme des pics de demande dans un système de chauffage urbain. Notre dataset consiste en quatre périodes séparées, avec 198 jours pour chaque période et 24 observations horaires dans chaque jour relatifs à la consommation de chaleur et le climat. Nous tenons en considération la nature fonctionnelle des données et proposons une méthodologie de prédiction basée sur la régression fonctionnelle. L'influence de variables explicatives exogènes est modelée d'une façon appropriée. Le résultats "out-of-sample" de l'approche proposée sont évalués.

Abstract

We consider the problem of short-term peak demand forecasting in a district heating system. Our dataset consists of four separated periods, with 198 days each period and 24 hourly observations within each day relative to heat consumption and climate. We take advantage of the functional nature of the data and we propose a forecasting methodology based on functional regression. The influence of exogenous explanatory variables is modelled in a suitable way. The out-of-sample performances of the proposed approach are evaluated.

Mots clés

Functional linear model, penalized splines estimation, peak load forecasting, district heating system

Introduction

Among the activities of support in the coordination, maintenance and planning of an energy system, the prediction of the load demand is one of the most important. In

particular, short-term forecasting, which is made within the 24 hours of the following day, and in special way the prevision of peaks of demand, plays a central role in guaranteeing an efficient generating capacity, maintaining the system stability.

In this work we consider the problem of modelling and predicting the peak of heat demand in a district heating (called also "teleheating") system. This consists in distributing the heat for residential and commercial requirements, via a network of insulated pipes. The dataset analyzed has been provided by AEM Turin Group, a municipal utility of the northern Italy city of Turin, which produces heat by means of cogeneration technology and distributes it, guaranteeing the heating to over a quarter of the town.

In the recent literature concerning load prediction in district-heating (see for example Dotzauer (2002) and Nielsen and Madsen (2006) for some applications and references), the algorithms employed are based on regression or time series models. They are often similar to the models used in the prediction of electrical-power loads (for a review see e.g. Weron (2006)). Weather factors are often used as major variables in predicting energy load and, among the others, the outdoor temperature is considered the most important factor in the short term forecasting.

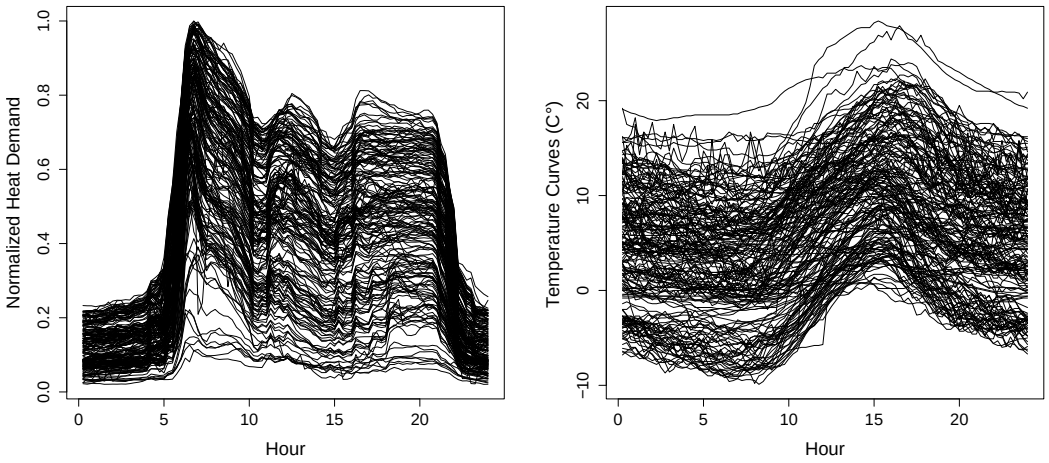
These methodologies skip sometimes the fact that the data used are discretization points of curves : they are observed with high frequency and are very highly correlated, exhibiting some seasonality patterns. This fact has stimulated us to explore the functional approach (see e.g. Ferraty and Vieu (2006) and Ramsay and Silverman (2006) for a review). We propose here a linear model, combining real regressors with functional ones.

Forecasting Peak Load

The dataset analyzed consists of measurements of heat consumption and temperature taken every hour, during the periods 15 October - 30 April in years 2001-02, 2002-03, 2003-04 and 2004-05. We take advantage of the functional nature of the data and we divide, in a natural way, the observed series of heat demand of each period in 198 *functional observations*, each one coincident with a specific daily load curve. We denote by $C_{y,d} = \{C_{y,d}(t),\ t \in [0,24]\}$ the daily load curve of period y and day d, with $y = 1,\ldots,4$ and $d = 1,2,\ldots,198$. Each of these functional data is observed on a finite mesh of discrete times : $t_1 < t_2 < \cdots < t_{24}$. Analogously we define the daily temperature curve. Figure 0.0.arabic@figure reports the observed loads and temperature curves of the first period.

Let us consider the forecasting problem of the daily peak load, defined as $P_{y,d} = \max_{j=1,\ldots,24} C_{y,d}(t_j)$. According to the literature (see e.g. Weron (2006)) we construct a linear model based on the decomposition of the load demand in a sum of two main components, namely the *load component* and the *weather-dependent component*, plus a stochastic residual. The first component includes :

FIG. 0.0.arabic@figure – Normalized load and temperature daily curves in the period 15 October 2001 - 30 April 2002.



- the seasonal effect, described by a suitable moving average of past daily means of consumptions ;
- the intra-daily effect, modelled by a weighed sum of second derivative of the load curve of the previous day. A reason to consider second derivative rather than the original curves is that data show an evident vertical shift and taking the second derivative annihilates this effect ;
- calendar effects (week-days, weekend-days, holy-days).

About the weather-dependent part, we use the daily temperature curve, weighted by a suitable functional coefficient.

Combining in an additive way the components previously identified and described, we arrive to the specification of a linear model with scalar response (the peak of heat demand), two scalar regressors (the seasonal part and the dummy indicating the calendar effects), and two functional regressors (the second derivative of the past daily load curve and temperature curve).

The model is estimated on the base of the training-set corresponding to the data observed in the first three periods (2001-02, 2002-03 and 2003-04) : we use here an estimation procedure proposed in Cardot *et al.* (2003) and based on B-splines. Then we carry out an out-of-sample forecasting study on the whole fourth period (2004-05), evaluating the results obtained. The estimated model fits well and the out-of-sample performances are good : we may compare them with the ones in Goia *et al.* (2010), where some functional and standard prediction methods are proposed to make forecasting on the same dataset.

Bibliographie

[1] Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model, *Statistica Sinica*, 13, 571-591.

[2] Dotzauer, E. (2002). Simple model for prediction of loads in district-heating systems. *Applied Energy*. 73(3), 277-284.

[3] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis*, Springer, New York.

[4] Goia, A., May, C. and Fusai, G. (2010). Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting*. Forthcoming.

[5] Nielsen, H.A., and Madsen, H. (2006). Modelling the heat consumption in district heating systems using a grey-box approach. *Energy and Buildings*, 38 :63-71.

[6] Ramsay, J.O. and Silverman, B.W. (2006). *Functional data analysis*, Second Edition, Springer, New York.

[7] Weron, R. (2006). *Modeling and Forecasting Electricity Loads and Prices : A Statistical Approach*, Wiley, Chichester.

Sélection de modèle incluant des composantes principales

Alois KNEIP et Pascal SARDA*

* Adresse pour correspondance :
Institut de Mathématiques, UMR 5219,
Équipe Statistique et Probabilités,
118, Route de Narbonne, 31062 Toulouse Cedex, France
e-mail : sarda@cict.fr

Résumé. Nous considérons un modèle de régression linéaire de grande dimension et plus précisément le cas d'un modèle factoriel pour lequel le vecteur des variables explicatives se décompose en la somme de deux termes aléatoires décrivant respectivement la variabilité spécifique et commune des prédicteurs. Nous montrons tout d'abord que les procédures de sélection de variables et d'estimation usuelles telles que le lasso ou le sélecteur Dantzig sont performantes dans ce contexte et sous l'hypothèse additionnelle que le vecteur des paramètres est *sparse*. Cette hypothèse peut être cependant restrictive. Nous introduisons ainsi un modèle de régression *augmenté* qui inclut les composantes principales. Nous montrons que ces composantes peuvent être convenablement estimées à partir de l'échantillon et nous nous concentrons ensuite sur les propriétés théoriques du modèle *augmenté*.

Abstract. We consider a high dimensional linear regression model and more precisely the case of a factor model where the vector of explanatory variables can be decomposed as a sum of two random terms representing respectively specific and common variability of the predictors. We show at first that usual prameter estimation and variable selection procedures such as Lasso or Dantzig selector are efficient in this context with the additional assumption that the vector of parameters is sparse. Such an assumption may be however restrictive. We thus introduce an augmented regression model which includes principal components. We show that these components can be accurately estimated from the sample and then we concentrate on the theoretical properties of the augmented model.

Mots clés. Modèle de régression linéaire, grande dimension, sélection de variables, composantes principales, Lasso, sélecteur Dantzig.

1. Introduction Dans de nombreuses applications le nombre de variables ou de paramètres est très élevé voire plus grand que la taille de l'échantillon. Une large littérature statistique est désormais consacrée à l'étude de problèmes en grande dimension. Un des modèles les plus souvent considérés est le modèle de régression linéaire :

$$Y_i \;=\; \boldsymbol{\beta}^T \mathbf{X}_i \;+\; \epsilon_i, \quad i = 1, \ldots, n, \tag{9}$$

où $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, sont des couples aléatoires avec $Y_i \in \mathbb{R}$ et $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$. Dans le modèle (9) $\boldsymbol{\beta}$ est un vecteur de paramètres dans $\mathbb{R}^p$ et $(\epsilon_i)_{i=1,\ldots,n}$ sont des v.a.r. i.i.d., indépendantes de $\mathbf{X}_i$, centrées et telles que $Var(\epsilon_i) = \sigma^2$. La dimension p du vecteur des paramètres est ici élevée comparativement à la taille de l'échantillon n.

Le modèle (9) décrit deux situations qui ont donné lieu à deux branches relativement indépendantes de la littérature statistique. La première correspond au cas où $\mathbf{X}_i$ représente un vecteur (de grande dimension) de différents prédicteurs alors qu'une autre situation apparaît lorsque les variables explicatives sont p points de discrétisation d'une même courbe. Dans ce dernier cas le modèle (9) est une version discrète du modèle linéaire fonctionnel. Pour chacune de ces situations des stratégies très différentes ont été adoptées afin d'estimer le vecteur des paramètres $\boldsymbol{\beta}$ et les hypothèses structurelles sous-jacents semblent être incompatibles.

Dans le premier cas les travaux reposent sur l'hypothèse que seul un nombre relativement petit de variables explicatives ont une influence significative sur la réponse Y_i ou qu'en d'autres termes le vecteur des coefficients β_j est *sparse* : $S := \#\{\beta_j | \beta_j \neq 0\} \ll p$. Cette hypothèse s'accompagne d'une condition sur les corrélations entre les différentes variables explicatives qui doivent être "suffisamment" faibles. Les procédures les plus populaires pour identifier et estimer les coefficients non nuls sont le *Lasso* et le *sélecteur Dantzig* : voir par exemple Tibshirani (1996), Bickel et al. (2009) et Candes and Tao (2007). Dans les travaux ayant trait à la statistique fonctionnelle on adopte des hypothèses très différentes. Si on considère le cas le plus simple où $X_{ij} = X_i(t_j)$ pour des fonctions aléatoires $X_i \in L^2([0,1])$ observées en des points équidistants $t_j = \frac{j}{p}$, on a alors $\beta_j := \frac{\beta(t_j)}{p}$, où $\beta(t) \in L^2([0,1])$ et lorsque $p \to \infty$, $\sum_j \beta_j X_{ij} = \sum_j \frac{\beta(t_j)}{p} X_i(t_j) \to \int_0^1 \beta(t) X_i(t) dt$. Par ailleurs, les corrélations entre les variables $X_{ij} = X_i(t_j)$ et $X_{il} = X_i(t_l)$, $j \neq l$, sont très fortes : lorsque $p \to \infty$, $corr(X_i(t_j), X_i(t_{j+m})) \to 1$ pour tout m fixé. Dans ce contexte, aucune variable $X_{ij} = X_i(t_j)$ n'a une influence particulière sur Y_i, et il y a une grand nombre de coefficients β_j qui sont proportionnels à $1/p$. Bien entendu, la réduction de dimension est également présente dans le cadre fonctionnel mais cependant elle s'obtient ici en réécrivant le modèle en termes d'une décompositon des prédicteurs sur une base "sparse", c'est-à-dire sur un petit nombre k de fonctions de base. Il est alors bien connu que la meilleure base possible au sens de l'erreur L^2 est celle fournie par les fonctions propres correspondant aux plus grande valeurs propres de l'opérateur de covariance de X_i. Parmi les nombreuses références sur le modèle linéaire fonctionnel citons Ramsay et

Dalzell (1981), Cardot et al. (1999), Cai et Hall (2007), Hall et Horowitz (2007), Cardot et al. (2007) et Crambes et al. (2009).

Le but de notre travail est de montrer qu'une combinaison des idées développées dans les deux approches ci-dessus conduit, pour le modèle "discret" (9), à une procédure d'estimation nouvelle qui peut être utile dans de nombreuses applications. Nous mous plaçons dans un cadre général dans lequel les variables explicatives peuvent provenir ou pas de la discrétisation d'une même courbe. Par ailleurs, nous considérons un *modèle factoriel* de la forme

$$\mathbf{X}_i \;=\; \mathbf{W}_i + \mathbf{Z}_i,\; i = 1, \ldots, n, \tag{10}$$

où $\mathbf{W}_i$ et $\mathbf{Z}_i$ sont deux vecteurs aléatoires indépendants de $\mathbb{R}^p$. Nous supposons que W_{ij} et Z_{ij} représentent des parties non négligeables de la variance de X_{ij} : chaque variable X_{ij}, $j = 1, \ldots, p$ possède une variabilité *spécifique* induite par Z_{ij} qui peut expliquer une partie de la réponse Y_i. D'un autre côté le terme W_{ij} représente une variabilité *commune* et les composantes principales, qui quantifient cette variabilité simultanée des régresseurs, peuvent également contribuer aux variations de la réponses. Ces arguments ont motivé l'utilisation d'un modèle de régression "augmenté" qui inclut les composantes principales comme variables explicatives additionnelles.

2. Le cadre de l'étude

Considérons le modèle de régression linéaire (9) avec des variables explicatives $\mathbf{X}_i$ qui peuvent être décomposées selon (10). On suppose de plus que $\mathbb{E}(X_{ij}) = 0$ pour tout $j = 1, \ldots, p$, et que

$$\sup_j \mathbb{E}(X_{ij}^2) \leq D_0 < \infty. \tag{11}$$

La matrice de variances-covariances de $\mathbf{\Sigma}$ de $\mathbf{X}_i$ se décompose sous la forme $\mathbf{\Sigma} \;=\; \mathbf{\Gamma} + \mathbf{\Psi}$, où $\mathbf{\Gamma} = \mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$, alors que $\mathbf{\Psi}$ est une matrice diagonale. On note dans la suite $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T$ la matrice de variances-covariances empirique basée sur l'échantillon $\mathbf{X}_i$, $i = 1, \ldots, n$.

Les variables indépendantes Z_{ij}, $j = 1, \ldots, p$, avec $var(Z_{ij}) = \sigma_j^2$, sont supposées vérifier la condition suivante : il existe deux constantes positives D_1 et D_2 telles que

$(A.1)$ $$0 \;<\; D_1 \;<\; \sigma_j^2 \;<\; D_2.$$

Nous supposons par ailleurs l'hypothèse suivante

$(A.2)$ Il existe $C_0 < \infty$ tel que les événements

$$\sup_{1 \leq j,l \leq p} |\frac{1}{n}\sum_{i=1}^n W_{ij} W_{il} - cov(W_{ij}, W_{il})| \leq C_0 \sqrt{\frac{\log p}{n}}, \tag{12}$$

$$\sup_{1\leq j,l\leq p} |\frac{1}{n}\sum_{i=1}^{n} Z_{ij}Z_{il} - cov(Z_{ij}, Z_{il})| \leq C_0\sqrt{\frac{\log p}{n}}, \tag{13}$$

$$\sup_{1\leq j,l\leq p} |\frac{1}{n}\sum_{i=1}^{n} Z_{ij}W_{il}| \leq C_0\sqrt{\frac{\log p}{n}}, \tag{14}$$

$$\sup_{1\leq j,l\leq p} |\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{il} - cov(X_{ij}, X_{il})| \leq C_0\sqrt{\frac{\log p}{n}}, \tag{15}$$

sont réalisés simulatanément avec la probabilité $A(n,p) > 0$, où $A(n,p) \to 1$ as $n, p \to \infty$, $\frac{\log p}{n} \to 0$.

On montre que si les composantes $\mathbf{W_i}$ et $\mathbf{Z}_i$ de $\mathbf{X}_i$ vérifient $\mathbf{W}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$ et $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$, alors $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma} + \mathbf{\Psi})$ et l'hypothèse (A.2) est alors satisfaite.

Concernant les variables W_{ij}, nous supposerons par ailleurs qu'un petit nombre de vecteurs propres de $\mathbf{\Gamma}$ suffit à bien approximer $\mathbf{W}_i$ (voir conditon (A.3) ci-dessous).

Envisageons pour le moment le cas d'un vecteur de paramètres $\boldsymbol{\beta}$ sparse :

$$\sharp\{\beta_j | \beta_j \neq 0\} \leq S \text{ for some } S \leq \tfrac{p}{2}.$$

Les procédures les plus populaires pour identifier et estimer les coefficients non nuls β_j sont Lasso et le sélecteur Dantzig. Dans un article récent Bickel et al. (2009) analysent ces méthodes. Ils donnent des conditions, *restricted eigenvalue assumptions*, portant sur les corrélations entre les variables X_{ij} et X_{il}, $j \neq l$, sous lesquelles ils obtiennent entre autres des bornes pour l'erreur L^q, $1 \leq q \leq 2$. On montre qu'une version de ces conditions, $RE(S, S, c_0)$, $c_0 = 1, 3$, est vérifiée par les variables explicatives ayant la structure (10) et lorsque (A.1) et (A.2) sont vérifiées. Notons que les variables X_{ij} sont préalablement normalisées de telle sorte que les éléments diagonaux de la matrice de variances-covariances empririque soient égaux à 1.

Nous avons remarqué plus haut que les variables W_{ij} peuvent également avoir une influence spécifique sur la réponse Y_i au travers d'un vecteur de paramètres non *sparse*. Dans ce cadre, nous pouvons intégrer les composantes principales aux variables explicatives. Nous présentons dans la section suivante le modèle augmenté résultant.

3. Le modèle augmenté

Nous notons dans la suite $\lambda_1 \geq \lambda_2 \geq \ldots$ les valeurs propres de $\frac{1}{p}\mathbf{\Gamma}$, $\mu_1 \geq \mu_2 \geq \ldots$, les valeurs propres de $\frac{\mathbf{1}}{\boldsymbol{p}}\mathbf{\Sigma}$ et $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq$ les valeurs propres de la matrice de variances-covariances $\frac{1}{p}\widehat{\mathbf{\Sigma}}$, alors que $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots$ et $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ sont des vecteurs propres orthonormés correspondant.

Le modèle incluant les composantes principales s'écrit

$$Y_i \;=\; \sum_{r=1}^{k} \alpha_r \xi_{ir} + \boldsymbol{\beta}^{*T}\mathbf{X}_i \;+\; \epsilon_i, \quad i = 1, \ldots, n, \tag{16}$$

où $\xi_{ir} = \boldsymbol{\delta}_r^T \mathbf{X}_i / \sqrt{p\mu_r}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T \in \mathbb{R}^k$ et $\boldsymbol{\beta}^* \in \mathbb{R}^p$ sont des vecteurs de paramètres. Nous supposons en outre que le vecteur $\boldsymbol{\beta}^*$ est sparse.

La première étape d'estimation du paramètres $(\alpha_1, \ldots, \alpha_k, \beta_1^*, \ldots, \beta_p^*)$ consiste à projeter le modèle à l'aide de la matrice de la projection sur l'espace engendré par les vecteurs propres correspondant aux k plus grandes valeurs propres de $\frac{1}{p}\widehat{\boldsymbol{\Sigma}}$

$$\widehat{\mathbf{P}}_k \;=\; \mathbf{I}_p - \sum_{r=1}^{k} \widehat{\boldsymbol{\psi}}_r \widehat{\boldsymbol{\psi}}_r^T.$$

Le modèle (16) s'écrit alors pour $i = 1, \ldots, n$

$$Y_i \;=\; \sum_{r=1}^{k} \alpha_r^* \widehat{\xi}_{ir} + \sum_{j=1}^{p} \beta_j^{**} \frac{(\widehat{\mathbf{P}}_k \mathbf{X}_i)_j}{\left(\frac{1}{n}\sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2\right)^{1/2}} \;+\; \epsilon_i^* + \epsilon_i, \tag{17}$$

où $\widehat{\xi}_{ir} = \widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i / \sqrt{p\widehat{\lambda}_r}$, $\alpha_r^* = \alpha_r + \sqrt{p\widehat{\lambda}_r} \sum_{j=1}^{p} \widehat{\psi}_{rj}\beta_j^*$, $\beta_j^{**} = \beta_j^* \left(\frac{1}{n}\sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2\right)^{1/2}$ and $\epsilon_i^* = \sum_{r=1}^{k} \alpha_r (\xi_{ir} - \widehat{\xi}_{ir})$.

Nous montrons tout d'abord dans la proposition ci-dessous que les valeurs et vecteurs propres théoriques sont bien approximés par leurs versions empiriques. Nous faisons les hypothèses additionnelles suivantes

(A.3) Il existe $1 \geq v(k) \geq 3C_0(\log p/n)^{1/2}$ tel que les valeurs propres de $\frac{1}{p}\boldsymbol{\Gamma}$ sont telles que

$$\min_{j,l \leq k, j \neq l} |\lambda_j - \lambda_l| \geq v(k), \quad \min_{j \leq k} \lambda_j \geq v(k).$$

Enfin nous supposons que n et p sont suffisamment grands pour que l'hypothèses suivante soit vérifiée

(A.4) $C_0(\log p/n)^{1/2} \geq \frac{D_0}{pv(k)}$.

Proposition 1 *Sous les hypothèses (A.2)-(A.4) et sous les événemnts (12) - (15) on a*

pour tout $r \leq k$ *et tout* $j = 1, \ldots, p$

$$|\lambda_r - \widehat{\lambda}_r| \leq \frac{D_2}{p} + C_0(\log p/n)^{1/2},\ |\mu_r - \widehat{\lambda}_r| \leq C_0(\log p/n)^{1/2} \tag{18}$$

$$\|\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r\|_2 \leq 5\frac{\frac{D_2}{p} + C_0(\log p/n)^{1/2}}{v(k)},\ \|\boldsymbol{\delta}_r - \widehat{\boldsymbol{\psi}}_r\|_2 \leq 3\frac{C_0(\log p/n)^{1/2}}{v(k)} \tag{19}$$

$$\psi_{rj}^2 \leq \frac{D_0 - D_1}{p\lambda_r} \leq \frac{D_0 - D_1}{pv(k)}, \tag{20}$$

$$\widehat{\psi}_{rj}^2 \leq \frac{D_0 + C_0(\log p/n)^{1/2}}{p\widehat{\lambda}_r} \leq 3\frac{D_0 + C_0(\log p/n)^{1/2}}{pv(k)}. \tag{21}$$

Nous sommes maintenant en mesure d'estimer les paramètres $\alpha*_r$ et β_j^{**} à l'aide du Lasso ou encore du sélecteur Dantzig. On en déduit ensuite des estimateurs de α_r et β_j^*. La condition $RE(k+S, k+S, c_0)$, $c_0 = 1, 3$, est satisfaite sous les hypothèses (A.2)-(A.4) ci-dessus. On en déduit alors, en utilisant les résultats de Bickel et al. (2009) des bornes pour la convergence L^q des estimateurs pour $1 \leq q \leq 2$.

Bibliographie

[1] Bickel, P.J., Ritov, Y. and Tsybakov, A. (2009) Simulataneous analysis of Lasso and Dantzig selector, *Ann. Statist.*, **37**, 1705-1732.

[2] Cai, T. and Hall, P. (2007). Prediction in functional linear regression, *Ann. Statisti.*, **34**, 2159-2179.

[3] Candès, E. and Tao, T. (2007). The Dantzig selector : statistical estimation when p is much larger than n, *Ann. Statist.*, **35**, 2013–2351, MR2382644.

[4] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model, *Statist. Prob. Letters*, **45**, 11-22.

[5] Cardot, H., Mas, A. and Sarda, P. (2007). CLT in functional linear regression models, *Prob. Theory Related Fields*, **138**, 325-361.

[6] Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing spline estimators for functional linear regression, *Ann. Statist.*, **37**, 35-72.

[7] Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression, *Ann. Statist.*, **35**, 70-91.

[8] Ramsay, J.O. and Dalzell, C.J. (1991). Some tools for functional data analysis (with discussion), *J. Roy. Statist. Soc. Ser B*, **53**, 539-572.

[9] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267-288.

Découpage de courbes de densité : Application au dépistage du cancer

Fabrice MORLAIS*, Frédéric FERRATY et Philippe VIEU

* Adresse pour correspondance :
ERI3 INSERM 'Cancers & Populations', EA 3936 Université Caen, Faculté de médecine - avenue côte de nacre - 14032 Caen cedex
&
Université Paul Sabatier, IMT UMR CNRS 5583, 118, Route de Narbonne, 31062 Toulouse Cedex, France

e-mail : fabricemorlais@yahoo.fr

Résumé

Le dépistage actuel du cancer broncho-pulmonaire est effectué à l'aide d'une radiographie pulmonaire, d'un scanner thoracique et d'un examen cytologique des expectorations. La cytologie automatisée des expectorations est une méthode permettant l'analyse informatique des cellules d'un crachat sur la lame d'un microscope. Comme une personne est représentée par l'ensemble des cellules de sa lame, il nous a paru intéressant d'utiliser la densité de probabilité comme unité statistique. La modélisation fonctionnelle des données, méthode pour laquelle l'unité statistique est à valeurs dans un espace infini, répond bien à cette problématique statistique puisque, par définition, une densité de probabilité est une fonction. Lors de cet exposé nous présenterons la méthode de classification supervisée de courbes de densité que nous avons développée, pour discriminer des personnes ayant un cancer et des personnes saines, et nous vous donnerons quelques résultats issus de données réelles.

Abstract

Screening of bronchopulmonary cancer is currently performed using chest-X ray, chest CT scan and cytological examination of expectorated sputum. Automated cytology of expectorated sputum is a method which enables sputum cells to be analyzed on a microscope slide. It seems interesting to use density probability as statistical unit because a person is represented by all the cells of her slide. As in functional data analysis statistical unit

takes values in an infinite dimensional space, we can use functional data analysis because density probability is a function. In this talk we will give the supervised classification of density curves that we have developed for discriminating persons having a cancer and control persons, and we will give some results that come from real data.

1. Introduction

En modélisation fonctionnelle (voir par exemple : Ferraty et Romain (2010), Ferraty et Vieu (2006) et Ramsay et Silverman (2005)) l'unité statistique n'est plus seulement représentée par un ensemble de n variables, à valeurs dans un espace $\mathbb{R}^n$, mais par une fonction, à valeurs dans un espace de dimensions infini. Une variable aléatoire χ est considérée comme fonctionnelle si elle prend des valeurs dans un espace infini, par exemple la variable fonctionnelle $\chi = \{X(t); t \in T\}$ avec $T \subset \mathbb{R}$ représente une courbe observée sur l'intervalle T de $\mathbb{R}$. La première étape à considérer dans une modélisation fonctionnelle est la transformation des données initialement discrétisées en une fonction continue. De nombreuses techniques statistiques existent pour réaliser cette transformation lorsque la fonction à estimer est une densité de probabilité (voir Wasserman (2007) pour la monographie la plus récente sur le sujet).
Bien que les méthodes statistiques classiques soient à peu près toutes développées dans le cadre fonctionnel, quelques difficultés surviennent pour les adapter lorsque les données fonctionnelles considérées sont des densités. Les méthodes statistiques listées ci-dessous sont des exemples permettant d'utiliser la densité de probabilité comme unité statistique. En statistique exploratoire, Kneip et Utikal (2001) ont développé une ACP fonctionnelle adaptée aux densités. En statistique supervisée, la méthode non paramétrique développée par Ferraty et Vieu (2003), dans le cadre de variables aléatoires fonctionnelles classiques, s'adapte aux densités puisqu'elle repose essentiellement sur la notion de distances entre courbes. Pour comparer globalement des courbes de densité, Delicado (2007) a développé une méthode modifiant l'ANOVA fonctionnelle de Cuevas *et al.* (2004) pour la rendre adaptable aux fonctions de densité de probabilité.
Nous avons privilégié l'approche de Ferraty et Vieu (2003) car elle correspondait parfaitement à notre problématique initiale qu'est la prévision. On trouvera dans le Chapitre 8 de Ferraty et Vieu (2006) et dans le Chapitre 10 de Ferraty et Romain (2010) des discussions bibliographiques plus complètes sur la classification de courbes. L'utilisation pratique de cette méthode nous a montré que, même dans le cadre de différences manifestes et visuelles entre groupes de densités, la discrimination au sens mathématique du terme n'était pas toujours retrouvée. Cela étant très probablement dû à certaines parties de la distribution perturbant l'analyse. Pour contourner ce problème nous avons développé une méthode statistique recherchant les morceaux de densité optimaux pour la discrimination. Cette méthode que nous nommerons par la suite 'Optimal cutting' sera présentée dans la partie 1. La partie 2 s'intéressera à la transformation de nos données initiales en fonctions,

en utilisant un estimateur à noyau standard. La partie 3 décrira les données réelles sur lesquelles ont été utilisées nos méthodes.

2. Optimal cutting

Soit $(\mathbf{X}_i, \mathbf{Y}_i)_{i=1,\ldots,n}$ un échantillon de paires indépendantes et identiquement distribuées de même loi $(\mathbf{X}, \mathbf{Y})$ à valeur dans $E \times \bar{G}$, avec $\mathbf{X}$ une variable aléatoire fonctionnelle définie sur un intervalle $[t_{min}, t_{max}] \subset \mathbb{R}$, (E, d) un espace vectoriel semi-métrique et $\mathbf{Y}$ une variable aléatoire catégorielle définie sur $\bar{G} = \{1, \ldots, G\}$. Soit t_0 un point de l'intervalle $[t_{min}, t_{max}]$. A partir de notre variable aléatoire fonctionnelle initiale $\mathbf{X}_i$, on construit deux nouvelles variables aléatoires fonctionnelles en découpant $\mathbf{X}_i$ de la façon suivante :

$$\mathbf{X}_i^{1,t_0} = \{\mathbf{X}_i(t), t \in [t_{min}, t_0[\}$$
$$\mathbf{X}_i^{2,t_0} = \{\mathbf{X}_i(t), t \in [t_0, t_{max}]\}$$

Nous disposons donc d'un n-échantillon de triplets indépendants $(\mathbf{X}_i^{1,t_0}, \mathbf{X}_i^{2,t_0}, \mathbf{Y}_i)$ avec $\mathbf{X}_i^{1,t_0}$ et $\mathbf{X}_i^{2,t_0}$ des variables aléatoires fonctionnelles prenant des valeurs dans un espace infini borné de $\mathbb{R}$ et $\mathbf{Y}$ une variable aléatoire catégorielle à valeurs dans $\bar{G} = \{1, \ldots, G\}$. A partir de ces morceaux de courbes indépendants, nous allons estimer pour chacun d'eux la probabilité a posteriori d'appartenance à G en fonction de t_0. Pour ce faire nous allons utiliser la classification supervisée non paramétrique de courbes fonctionnelles développée par Ferraty et Vieu (2003). Nous avons découpé notre échantillon initial en deux échantillons : un échantillon d'apprentissage (L) et un échantillon test (T). L'estimation de ces probabilités a posteriori se fera, pour chaque morceau de courbe X, de la façon suivante :

$$p_g(X) = P(\mathbf{Y} = g | \mathbf{X} = X), g \in \bar{G}$$

où $\mathbf{X}$ est une variable aléatoire fonctionnelle et X est une réalisation de cette variable aléatoire fonctionnelle. Une fois les G probabilités a posteriori estimées, nous affecterons à $\hat{Y}(x)$ le numéro de groupe de plus forte probabilité (classifieur bayésien) :

$$\hat{Y}(X) = \arg\max_{g \in \bar{G}} \hat{p}_g(X)$$

Avant de définir notre estimateur à noyau de la probabilité a posteriori, remarquons que :

$$p_g(X) = P(\mathbf{Y} = g | \mathbf{X} = X) = E(\mathbb{1}_{[\mathbf{Y}=g]} | \mathbf{X} = X)$$

Cette probabilité peut donc être estimée en terme d'espérance conditionnelle et nous pouvons donc utiliser un estimateur de type noyau pour prédire cette espérance conditionnelle :

$$\hat{p}_g(X) = \hat{p}_{g,h}(X) = \frac{\sum_{i \in L} \mathbb{1}_{[\mathbf{Y}_i = g]} K\left(\frac{d(\mathbf{X}_i, X)}{h}\right)}{\sum_{i \in L} K\left(\frac{d(\mathbf{X}_i, X)}{h}\right)}$$

où K est un noyau asymétrique, h est la taille de fenêtre du noyau et d est une semi-métrique. L'expression ci-dessus correspond à la régression d'une variable dichotomique sur une variable fonctionnelle. Le choix de la taille de fenêtre h optimale se fera en minimisant une fonction de coût du type :

$$h_L^{opt} = \arg\inf_h Loss_L(h)$$
$$Loss_L(h) = \sum_{j\in L}\sum_{g\in G}\left(\hat{p}_{g,h} - \mathbb{1}_{[Y_j=g]}\right)^2$$

Pour chaque valeur $t_0 = \{t_{min}, t_{min+1}, \ldots, t_{max},\}$ et pour chaque morceau de courbe X^1 et X^2, nous obtenons une taille de fenêtre minimisant la fonction $Loss$. Nous identifions ensuite la valeur t^{opt} et le morceau de courbe optimal pour lesquels la fonction $Loss$ atteint son minimum.
Ayant un effectif faible nous avons évalué la qualité de prédiction de la modélisation par validation croisée (Hatsie et al. (2009)) :

$$Misclass = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[Y_i\neq\hat{Y}_i]}$$

où $\hat{Y}_i$ est la prédiction obtenue sur l'échantillon $L_i = L_{\{j,j\neq i\}}$ à partir du morceau de courbe optimal précédemment défini et Y_i est une réalisation de la variable aléatoire $\mathbf{Y}_i$.

3. Estimation de densité

Les variables fonctionnelles $\mathbf{X}_i$ sont des densités de probabilité provenant d'un échantillon de variables aléatoires indépendantes et identiquement distribuées $\mathbf{Z}_{i,1}, \ldots, \mathbf{Z}_{i,n_i}$. Pour estimer les densités de probabilité, nous avons utilisé un estimateur non paramétrique à noyau.

$$\hat{X}_i(t) = \frac{1}{n_i h_{d,i}}\sum_{j=1}^{n_i} K_d\left(\frac{t-\mathbf{Z}_{i,j}}{h_{d,i}}\right)$$

où K_d est un noyau symétrique tel que :

$$\int_{-\infty}^{+\infty} K_d(t)dt = 1$$

et $h_{d,i}$ la taille de la fenêtre. Deux procédures permettant la sélection automatique des fenêtres ont été implémentées. La première est standard et utilise la méthode 'Plug-in' de Sheather et Jones (1991). La seconde opère un choix adaptatif de la fenêtre lié au problème de discrimination. Du point de vue de la méthodologie, il suffit d'appliquer le découpage optimal aux densités obtenues avec le choix de fenêtre Plug-in. Pour le second

choix de fenêtre, nous avons modifié l'optimal cutting. D'une part on simplifie le problème en prenant des fenêtres telles que :

$$h_{d,i} = h_{0,i} \times \alpha$$

où $h_{0,i}$ est la fenêtre obtenue par la la méthode Plug-in de Sheather et Jones (1991) et α est une constante. D'autre part, on sélectionne la taille de fenêtre $h_{d,i}$ ayant la fonction de coût *Loss* minimale. De plus, nous avons raffiné notre méthode en ajoutant un prétraitement à nos données. L'enregistrement de courbes (Kneip et Engel (1995), Kneip et Gasser (1992), Ramsay et Silverman (2005)) est utile lorsqu'une forme commune semble apparaître, et lorsque les réalisations individuelles de cette forme diffèrent en phase (variation horizontale) ou en amplitude (variation verticale). La non prise en compte de ces deux phénomènes lors de la modélisation peut amener à prendre en compte des déformations pouvant dégrader la qualité de prédiction. Dans notre étude nous ne nous sommes pas intéressés aux variations d'intensités (ou variation en amplitude, ou variation verticale) car nous souhaitions garder la caractéristique de densité de nos fonctions $\int_{t_{min}}^{t_{max}} X(t)dx = t$. Nous avons simplement aligné nos densités par rapport au mode principal.

4. Présentation des données réelles

Le dépistage actuel du cancer broncho-pulmonaire peut être effectué à l'aide d'une radiographie pulmonaire, d'un scanner thoracique et d'un examen cytologique des expectorations. En radiographie pulmonaire et en scanner thoracique, le praticien s'intéresse à la présence et à l'évolution de la taille des nodules (regroupement de cellules) suspects de cancer. En cytologie 'conventionnelle' des expectorations le pathologiste s'intéresse à la présence de cellules cancéreuses dans un crachat à l'aide d'un microscope optique. Quelques récents travaux (Belien *et al.* (1997), Doudkine *et al.* (1995), Palcic *et al.* (2002) et Payne *et al.* (1997)) ont montré l'intérêt d'une nouvelle technique cytologique des expectorations dans le dépistage précoce de cancers : la cytologie automatisée. La cytologie automatisée des expectorations est une méthode permettant l'analyse informatique des cellules d'un crachat sur la lame d'un microscope. Une caméra numérique reliée à un ordinateur découpe l'image de cette lame en petites images qui sont alors stockées dans l'ordinateur. Ces images sont ensuite traitées par un logiciel d'imagerie qui détecte les cellules du prélèvement, par une méthode de détection de contours, et qui les analyse. Ainsi pour chaque cellule, un certain nombre de paramètres de forme, de texture et d'intensité sont mesurés. Il est important de remarquer que même dans les cas de cancers, la grande majorité des cellules d'une lame est normale. Les cas de cancers présentent généralement très peu de cellules suspectes (ou malignes).
Nous avons comparé les distributions des cellules des individus sains et des individus ayant un cancer pour essayer de déterminer les caractéristiques ou groupes de caractéristiques cellulaires discriminants le mieux ces deux populations. Les résultats de notre méthode

seront donnés sur ce jeu de données, puis comparés à des méthodes de classification supervisée classiques.

Bibliographie

Belien, J.A.M., Baak, J.P.A., van Diest, P.J., Misere, B.N.L.H.M., Meijer, G.A., Bergers, L. (1997) Prognostic value of image and flow cytometric DNA ploidy assessments in invasive breast cancer, *Electr. J. Pathol*, 3, 972-979

Cuevas, A., Febrero, M. et Fraiman, R. (2004) An anova test for functional data, *Computational Statistics and Data Analysis*, 44, 111–122.

Delicado, P. (2007) Functional k-sample problem when data are density functions, *Computational Statistics and Data Analysis*, 22, 391–440.

Doudkine, A., MacAulay, C., Poulin, N., Palcic, B. (1995) Nuclear texture measurements in image cytometry, *Pathologica*, 87, 286-299

Ferraty, F., Vieu, P. (2003) Curves discrimination : a non parametric functional approach, *Computational Statistics and Data Analysis*, 44, 161–173.

Ferraty, F., Vieu P. (2006) Nonparametric Functional Data Analysis, Springer.

Ferraty, F. Romain, Y. (2010). Handbook on functional data analysis and related topics. Oxford University Press, to appear.

Hastie, T., Tibshirani, E., Friedman, J. (2009) The Elements of Statistical Learning, Springer

Kneip, A., Gasser, T. (1992) Statistical tools to analyse data representing a sample of curves, *The Annals of Statistics*, 20, 1266-1305

Kneip, A., Engel, J. (1995) Model estimation in nonlinear regression under shape invariance, *The Annals of Statistics*, 23, 551-570

Kneip, A., Utikal, K.J. (2001) Inference for densities families unsing functional principal component analysis, *Journal of the American Statistical Association*, 96, 519–542.

Palcic, B., Garner, D.M., Beveridge, J., xiao Rong Sun, Doudkine, A., Macaulay, C., Lam, S.,Payne, P.W. (2002) Increase of sensitivity of sputum cytology using high-resolution image cytometry : field study results,*Cytometry*, 50, 168-176

Payne, P.W., Sbebo, T.J., Doudkine, A., Garner, D., MacAulay, C., Lam, S., LeRichie, J.C., Palcic, B. (1997), Sputum screening by quantitative microscopy : a reexamination of a portion of the National Cancer Institute Cooperative Early Lung Cancer Study, *Mayo Clin Proc*, 72, 697-704

Ramsay, J.O., Silverman, B.W. (2005) Functional Data Analysis, Springer.

Sheather, S.J., Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society. Series B (Methodological)*, 53, 983-990

Wasserman, L. (2007) All of Nonparametric Statistics, Springer.

Functional Common Principal Components Models

Graciela BOENTE, Daniela RODRIGUEZ et Mariela SUED*

* Adresse pour correspondance :
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires and CONICET
Argentina

e-mail : msued@dm.uba.ar

Abstract

In this paper, we discuss the extension to the functional setting of common principal component model that has been widely studied when dealing with multivariate observations. We provide estimators of the common eigenfunctions and to study their asymptotic behavior.

Résumé

Dans cet exposé, nous discutons l'extension au cas fonctionnel du modèle de composantes principales communes, qui a été abondamment étudié lorsqu'on s'intéresse à des observations multivariées. Nous considérons des estimateur pour les composantes principales communes dont la distribution asymptotique est étudiée.

1. Introduction

Functional data analysis is an emerging field in statistics that has received considerable attention during the last decade due to its applications to different fields. It provides modern data analytical tools for data that are recoded as a continuous phenomenon over a period of time. Because of the intrinsic nature of these data, they can be viewed as realizations of random functions $X_1(t), \ldots, X_n(t)$ often assumed to be in $L^2([0,1])$. In this context, principal components analysis offers an effective way for dimension reduction and it has been extended from the traditional multivariate setting to accommodate functional data. In the functional data analysis literature, it is usually referred to as functional principal component analysis (FPCA).

In many situations, we have independent observations $X_{i,1}(t), \cdots, X_{i,n_i}(t)$ from k independent samples of random functions in $L^2[0,1]$ with mean μ_i and different covariance

operators $\mathbf{\Gamma}_i$. However, as it is the case in the finite–dimensional setting, the covariance operators may exhibit some common structure. The common principal components model, introduced by Flury [**?**] for $p-$th dimensional data, generalizes proportionality of the covariance matrices by allowing the matrices to have different eigenvalues but identical eigenvectors. A natural extension to the functional setting of the common principal components model is to assume that the covariance operators $\mathbf{\Gamma}_i$ have common eigenfunctions $\phi_j(t)$ but different eigenvalues λ_{ij}. We will denote this model the functional common principal component (FCPC) model.

The aim of this work is to provide estimators of the common eigenfunctions under a FCPC model and to study their asymptotic behavior. Proofs are given by Boente, Rodriguez and Sued [**?**].

2. Notation and Preliminaries

Let $X_{i,1}(t), \cdots, X_{i,n_i}(t)$, $1 \leq i \leq k$, be independent observations from k independent samples of smooth random functions in $L^2\mathcal{I}$, where $\mathcal{I} = [0,1]$, with mean μ_i. Denote by γ_i and $\mathbf{\Gamma}_i$ the covariance function and operator, respectively, related to each population. To be more precise, we are assuming that $\{X_{i,1}(t) : t \in \mathcal{I}\}$ are k stochastic processes defined in $(\Omega, \mathcal{A}, P)$ with continuous trajectories, mean μ_i and finite second moment, i.e., $E\left(X_{i,1}(t)\right) = \mu_i(t)$ and $E\left(X_{i,1}^2(t)\right) < \infty$ for $t \in \mathcal{I}$. Each covariance function $\gamma_i(t,s) = \text{COV}(X_{i,1}(s), X_{i,1}(t))$, $s,t \in \mathcal{I}$ has an associated linear operator $\mathbf{\Gamma}_i : L^2[0,1] \to L^2[0,1]$ defined as $\left(\mathbf{\Gamma}_i\, u\right)(t) = \int_0^1 \gamma_i(t,s)u(s)ds$, for all $u \in L^2[0,1]$. As in the case of one population, throughout this paper, we will assume that the covariance operators satisfy $\|\gamma_i\|^2 = \int_0^1 \int_0^1 \gamma_i^2(t,s)dtds < \infty$. Therefore, $\mathbf{\Gamma}_i$ is a self–adjoint continuous linear operator. Moreover, $\mathbf{\Gamma}_i$ is a Hilbert-Schmidt operator. The FCPC model assume that the covariance operators $\mathbf{\Gamma}_i$ have common eigenfunctions $\phi_j(t)$, to be estimated.

When dealing with one population, estimators of the eigenfunctions and eigenvalues of $\mathbf{\Gamma}$ were considered by Dauxois, Pousse and Romain [**?**], in a natural way through the empirical covariance operator. In the present setting, we will give two proposals to estimate the common eigenfunctions under a FCPC model. Both of them are based on estimators $\widehat{\mathbf{\Gamma}}_i$ of the covariance operators $\mathbf{\Gamma}_{i,\mathrm{R}}$, like $\widehat{\mathbf{\Gamma}}_{i,\mathrm{R}}$, the operator associated to the empirical covariance functions $\widehat{\gamma}_{i,\mathrm{R}}(s,t) = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(X_{i,j}(s) - \overline{X}_i(s)\right)\left(X_{i,j}(t) - \overline{X}_i(t)\right)$.

Assume $n_i = \tau_i N$ with $0 < \tau_i < 1$ fixed numbers such that $\sum_{i=1}^k \tau_i = 1$ and where $N = \sum_{i=1}^k n_i$ denotes the total number of observations in the sample. Define the weighted covariance function as $\gamma = \sum_{i=1}^k \tau_i\gamma_i$ and its related operator as $\mathbf{\Gamma} = \sum_{i=1}^k \tau_i\mathbf{\Gamma}_i$. Therefore, $\widehat{\gamma}_{\mathrm{R}} = \sum_{i=1}^k \tau_i\widehat{\gamma}_{i,\mathrm{R}}$ and $\widehat{\mathbf{\Gamma}}_{\mathrm{R}} = \sum_{i=1}^k \tau_i\widehat{\mathbf{\Gamma}}_{i,\mathrm{R}}$ provide estimators of γ and $\mathbf{\Gamma}$, respectively. It is worth noticing that our results do not make use of the explicit expression of the covariance operator estimators, but they only require their consistency and asymptotic normality.

3. The proposals

Let us assume that the FCPC model hold, i.e., $\mathbf{\Gamma}_i$ have common eigenfunctions $\phi_j(t)$ but possible different eigenvalues λ_{ij}, where $\lambda_{ij} = \langle \phi_j, \mathbf{\Gamma}_i \phi_j \rangle$. Moreover, throughout this paper we will assume that

A1. $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \geq \lambda_{ip} \geq \lambda_{ip+1} \cdots$, for $1 \leq i \leq k$

A2. There exists ℓ such that for any $1 \leq j \leq \ell$, there exists $1 \leq i \leq k$ such that $\lambda_{ij} > \lambda_{i\,j+1}$.

The first proposal is based on the fact that under the FCPC model, the common eigenfunctions $\{\phi_j : j \geq 1\}$ are also a basis of eigenfunctions for the operator $\mathbf{\Gamma} = \sum_{i=1}^k \tau_i \mathbf{\Gamma}_i$, with eigenvalues given by $\nu_1 = \sum_{i=1}^k \tau_i \lambda_{i1} \geq \cdots \geq \nu_p = \sum_{i=1}^k \tau_i \lambda_{ip} \geq \nu_{p+1} = \sum_{i=1}^k \tau_i \lambda_{i\,p+1} \cdots$. Note that **A1** and **A2** entail that the first ℓ eigenfunctions will be related to the ℓ largest eigenvalues of the operator $\mathbf{\Gamma}$, having multiplicity one and being strictly positive. A first attempt to estimate the common eigenfunctions consists in considering the eigenfunctions $\widetilde{\phi}_j$ related to the largest eigenvalues $\widehat{\nu}_j$ of a consistent estimator $\widehat{\mathbf{\Gamma}}$ of $\mathbf{\Gamma}$, obtained as $\widehat{\mathbf{\Gamma}} = \sum_{i=1}^k \tau_i \widehat{\mathbf{\Gamma}}_i$ where $\widehat{\mathbf{\Gamma}}_i$ denotes any estimator of the $i-$th covariance operator. Example of such estimators are the associated to the empirical covariance functions $\widehat{\gamma}_{i,\mathrm{R}}$. The eigenvalue estimators can then be defined as $\widehat{\lambda}_{ij} = \langle \widetilde{\phi}_j, \widehat{\mathbf{\Gamma}}_i \widetilde{\phi}_j \rangle$.

The second proposal tries to improve the efficiency of the previous one for gaussian processes. To that purpose, we will have in mind that, in the finite–dimensional case, the maximum likelihood estimators of the common directions for normal data solve a system of equations involving both the eigenvalue and eigenvector estimators (see Flury, [**?**]). Using consistent estimators of the eigenvalues, we generalize the system obtained by Flury to the infinite–dimensional case. Effectively, let $\widehat{\lambda}_{ij}$ be initial estimators of the eigenvalues and $\widehat{\mathbf{\Gamma}}_i$ any consistent estimator of the covariance operator of the $i-$th population. Define for $j \leq \ell$ and $m \leq \ell$, $\widehat{\mathbf{\Gamma}}_{mj} = \sum_{i=1}^k \tau_i \frac{\widehat{\lambda}_{ij} - \widehat{\lambda}_{im}}{\widehat{\lambda}_{im}\widehat{\lambda}_{ij}} \widehat{\mathbf{\Gamma}}_i$, which will be asymptotically well defined under **A2** if in addition $\lambda_{i\ell} > 0$ for $1 \leq i \leq k$. Let us consider the solution $\widehat{\phi}_j$ of the system of equations

$$\begin{cases} \delta_{mj} = & \langle \widehat{\phi}_m, \widehat{\phi}_j \rangle \\ 0 = & \langle \widehat{\phi}_m, \widehat{\mathbf{\Gamma}}_{mj} \widehat{\phi}_j \rangle \qquad 1 \leq j < m \, . \end{cases} \tag{22}$$

4. Asymptotic distribution

It is clear that consistency of each population covariance operator estimator ensures consistency of the pooled one. The results in Section 2.1 of Dauxois, Pousse and Romain [**?**], allow to obtain the asymptotic distribution of the estimators of the common eigenfunctions. In particular, we obtain the following result (see, Boente, Rodriguez and Sued, [**?**], for details).

Proposition 4.1. *Let us assume that $\widehat{\mathbf{\Gamma}}_i$ is the empirical operator, $\widehat{\mathbf{\Gamma}}_{i,\mathrm{R}}$, that $E(\|X_{i,1}\|^4) < \infty$, for $1 \leq i \leq k$, and that **A1** and **A2** hold. For each eigenfunction ϕ_j of $\mathbf{\Gamma}$ related to*

the eigenvalue $\nu_j = \sum_{i=1}^k \tau_i \lambda_{ij}$ *with multiplicity one, we have that*

a) $\sqrt{N}(\widetilde{\phi}_j - \phi_j, \phi_j) \xrightarrow{p} 0$

b) For any $j \neq m$ $\sqrt{N}\langle \widetilde{\phi}_j - \phi_j, \phi_m \rangle \rightarrow \mathcal{N}(0, \sigma^2_{mj})$ *with*

$$\sigma^2_{jm} = \left\{ \sum_{i=1}^k \tau_i (\lambda_{ij} - \lambda_{im}) \right\}^{-2} \sum_{i=1}^k \tau_i \lambda_{im} \lambda_{ij} E[f^2_{im} f^2_{ij}]$$

Moreover, if $X_{i,1}$ *are gaussian processes, for all* $1 \leq i \leq k$*, we get that*

$$\sigma^2_{jm} = \left\{ \sum_{i=1}^k \tau_i (\lambda_{ij} - \lambda_{im}) \right\}^{-2} \sum_{i=1}^k \tau_i \lambda_{im} \lambda_{ij} \, . \tag{23}$$

The following Theorem provides the asymptotic behavior of the eigenvalue estimators under mild conditions on the eigenfunction estimators. It can be used to derive the asymptotic normality of the eigenvalue estimators when using, either Proposal 1 or Proposal 2 to estimate the eigenfunctions.

Theorem 4.1. *Let* $\widehat{\mathbf{\Gamma}}_i$ *be an estimator of the covariance operator of the* $i-th$ *population such that* $\sqrt{n_i}(\widehat{\mathbf{\Gamma}}_i - \mathbf{\Gamma}_i) \xrightarrow{\mathcal{D}} \mathbf{U}_i$*, where* $\mathbf{U}_i$ *is zero mean gaussian random element with covariance operator* $\mathbf{\Upsilon}_i$*. Let* $\widetilde{\phi}_j$ *be consistent estimators of the common eigenfunctions such that* $\sqrt{N}\left(\widetilde{\phi}_j - \phi_j\right) = O_p(1)$ *and define estimators of* λ_{ij} *as* $\widehat{\lambda}_{ij} = \langle \widetilde{\phi}_j, \widehat{\mathbf{\Gamma}}_i \widetilde{\phi}_j \rangle$*. For any fixed* m*, denote* $\widehat{\mathbf{\Lambda}}_i^{(m)} = \left\{ \sqrt{n_i}\left(\widehat{\lambda}_{ij} - \lambda_{ij}\right) \right\}_{1 \leq j \leq m}$*. Then,*

a) For each $1 \leq i \leq k$*,* $\sqrt{n_i}\left(\widehat{\lambda}_{ij} - \lambda_{ij}\right)$ *has the same asymptotic distribution as* $\sqrt{n_i}\left(\langle \phi_j, \widehat{\mathbf{\Gamma}}_i \phi_j \rangle - \lambda_{ij}\right)$*.*

b) For any m *fixed,* $\widehat{\mathbf{\Lambda}}_1^{(m)}, \ldots, \widehat{\mathbf{\Lambda}}_k^{(m)}$ *are asymptotically independent.*

c) If, in addition, the covariance operator $\mathbf{\Upsilon}_i$ *of* $\mathbf{U}_i$ *is given by*

$$\mathbf{\Upsilon}_i = \sum_{m,r,o,p} s_{im} s_{ir} s_{io} s_{ip} E[f_{im} f_{ir} f_{io} f_{ip}] \, \phi_m \otimes \phi_r \tilde{\otimes} \phi_o \otimes \phi_p - \sum_{m,r} \lambda_{im} \lambda_{ir} \, \phi_m \otimes \phi_m \tilde{\otimes} \phi_r \otimes \phi_r$$

then, $\widehat{\mathbf{\Lambda}}_i^{(m)}$ *is jointly asymptotically normally distributed with zero mean and covariance matrix* $\mathbf{C}^{(i,m)}$ *such that* $\mathbf{C}^{(i,m)}_{jj} = \lambda^2_{ij}\left[E\left(f^4_{ij}\right) - 1\right]$ *and* $\mathbf{C}^{(i,m)}_{js} = \lambda_{ij}\lambda_{is}\left[E\left(f^2_{ij} f^2_{is}\right) - 1\right]$*, that is, the asymptotic variance of* $\sqrt{n_i}\left(\widehat{\lambda}_{ij} - \lambda_{ij}\right)$ *is given by* $\lambda^2_{ij}\left[E\left(f^4_{ij}\right) - 1\right]$ *and the asymptotic correlations are given by*

$$\frac{E\left(f^2_{ij} f^2_{is}\right) - 1}{\left[E\left(f^4_{ij}\right) - 1\right]^{\frac{1}{2}} \left[E\left(f^4_{is}\right) - 1\right]^{\frac{1}{2}}} \, .$$

Moreover, in the normal case, we get that the components of $\widehat{\mathbf{\Lambda}}_i^{(m)}$ *are asymptotically independent with asymptotic variances* $2\lambda_{ij}^2$.

In order to study the asymptotic behavior of the second proposal, let $\mathbf{\Gamma}_{mj} = \sum_{i=1}^{k} \tau_i \left[(\lambda_{ij} - \lambda_{im}) / (\lambda_{im}\lambda_{ij}) \right] \mathbf{\Gamma}_i$ and denote $\phi_j^\star$ any solution of

$$\left\{ \begin{array}{rcll} \delta_{mj} = & \langle \phi_m^\star, \phi_j^\star \rangle & & \\ 0 = & \langle \phi_m^\star, \mathbf{\Gamma}_{mj} \phi_j^\star \rangle & \qquad 1 \leq j < m \ . \end{array} \right. \tag{24}$$

It is easy to see that if the covariance operators satisfy a FCPC model, then ϕ_j satisfies (24). Moreover, in Boente, Rodriguez and Sued [**?**] the consistency of the estimators defined through (22) is derived under mild conditions. The following result states the asymptotic behavior of the coordinates $\{\langle \widehat{\phi}_j, \phi_s \rangle : s \geq 1\}$ of the common eigenfunctions estimators $\widehat{\phi}_j$ defined through Proposal 2 that will allow to establish an improvement in efficiency for gaussian processes.

Theorem 4.1. *Let* $\widehat{\mathbf{\Gamma}}_i$ *be an estimator of the covariance operator of the* $i-th$ *satisfying the same hipotheses as in Theorem 4.1. Let* $\widehat{\lambda}_{ij}$ *be consistent estimators of the eigenvalues of the* $i-th$ *population* λ_{ij} *and* $\widehat{\phi}_j$ *consistent estimators of the common eigenfunctions* ϕ_j*, solution of (22) and denote* $\widehat{g}_j = \sqrt{N}\left(\widehat{\phi}_j - \phi_j\right)$*. Assume* ***A1****,* ***A2*** *and that* $\lambda_{i\ell} > 0$*, for all* $1 \leq i \leq k$*. If, in addition, for any* $j \leq \ell$*,* $m \leq \ell$*, the following two conditions hold*

i) $\langle \widehat{g}_j, \widehat{\phi}_m - \phi_m \rangle = o_p(1)$

ii) the operators $\mathbf{\Gamma}_i$ *have finite rank* ℓ*, for all* $1 \leq i \leq k$*, or* $\langle \widehat{g}_j, \mathbf{\Gamma}_i \left(\widehat{\phi}_m - \phi_m\right) \rangle = o_p(1)$*.*

then, for any $j \leq \ell$*,* $m \leq \ell$*,* $m \neq j$ *we have that*

a) $\langle \widehat{g}_m, \phi_j \rangle$ *has the same asymptotic distribution as* $-\langle \widehat{g}_j, \phi_m \rangle$*.*

b) For $j < m$*,* $\langle \widehat{g}_j, \phi_m \rangle \xrightarrow{\mathcal{D}} \mathcal{N}(0, \theta_{jm}^2)$*, where*

$$\theta_{jm}^2 = \frac{\displaystyle\sum_{i=1}^{k} \tau_i \frac{(\lambda_{im} - \lambda_{ij})^2}{\lambda_{im}\lambda_{ij}} E\left(f_{im}^2 f_{ij}^2\right)}{\left\{ \displaystyle\sum_{i=1}^{k} \tau_i \frac{(\lambda_{im} - \lambda_{ij})^2}{\lambda_{im}\lambda_{ij}} \right\}^2} \ . \tag{25}$$

Remark 4.1. Note that in the gaussian case, we get $E\left(f_{im}^2 f_{ij}^2\right) = 1$ and so the asymptotic variance of coordinates of the common eigenfunction estimates, defined through Proposal 2, reduces to

$$\theta_{jm}^2 = \left\{ \sum_{i=1}^{k} \tau_i \frac{(\lambda_{im} - \lambda_{ij})^2}{\lambda_{im}\lambda_{ij}} \right\}^{-1}$$

On the other hand, the common eigenfunction estimates, defined through Proposal 1, have asymptotic variances σ_{jm}^2 given by (23). Since $\theta_{jm}^2 \leq \sigma_{jm}^2$, we obtain that the estimates of Proposal 2 are more efficient that those of Proposal 1 for gaussian processes.

Bibliographie

Boente, G. ; Rodriguez, D. and Sued, M. (2009). Inference under functional proportional and common principal components models. Available at
http ://www.ic.fcen.uba.ar/preprints/boente_rodrigues_sued.pdf

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function : Some applications to statistical inference. *J. Multivar. Anal.*, **12**, 136-154.

Flury, B. K. (1984). Common principal components in k groups. *J. Am. Statist. Assoc.* **79,** 892-8.