Publications du Groupe de travail STAPH: Statistique Fonctionnelle et Opératorielle

STAPH-2011-01 *Recueil de résumés de l'année 2010-2011*

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

Institut de Mathématiques, Université Paul Sabatier, Toulouse, France

.

STAPH: Groupe de travail en Statistique Fonctionnelle et Opératorielle. Présentation des activités 2009-2010.

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

Adresse pour correspondance:
Institut de Mathématiques, Université Paul sabatier
Toulouse

Dans ce document, sont regroupés les résumés des exposés donnés lors des séances du groupe de travail en Statistique Fonctionnelle et Opératorielle STAPH au cours de l'année 2010-2011. Nous tenons tout d'abord à remercier chaleureusement tous les orateurs pour la qualité de leurs contributions qui ont donné lieu à des échanges fructueux. Nous saluons tout particulièrement les nouveaux venus à notre groupe de travail (la moitié des orateurs) et avec lesquels des membres de STAPH ont noué des liens scientifiques ou amorcé des collaborations. De manière générale, nous avons accueilli des doctorants, des jeunes chercheurs ou encore des chercheurs confirmés, suivant en cela un objectif de pluralité qui nous tient à coeur.

Cette introduction est également l'occasion pour nous de rappeler que l'activité de notre groupe, au cours de cette douzième année d'existence, s'est essentiellement centrée sur l'organisation du second meeting international en Statistique Fonctionnelle et Opératorielle (IWFOS2011) qui s'est tenu en Juin à Santander en Espagne sous la co-présidence franco-espagnole de Juan Cuesta-Albertos (Santander), Frédéric Ferraty (Toulouse) et Wenceslao Gonzalez-Manteiga (Saint Jacques de Compostelle) et qui a connu un franc succès, en regroupant une centaine de participants venus des cinq continents autour d'orateurs choisis parmi les plus actifs sur la scène internationale en Statistique Fonctionnelle. L'ouvrage [1] synthétise les contributions à cette conférence. On pourra se rendre compte à la lecture de ces résumés que les exposés qui y sont présentés témoignent du dynamisme et de la richesse de la recherche en Statistique Fonctionnelle. On y note une diversité et un renouvellement des thèmes abordés qui touchent aux différents aspects de la Statistique Fonctionnelle : modèles nonparamétriques et semiparamétriques fonctionnels, statistique opératorielle, modèles pour variables fonctionnelles, sélection de variable ... Les différentes approches mêlent recherche théorique et applications (économétrie, océanologie, chimiométrie, ...) : c'est un aspect

auquel nous tenons particulièrement.

Pour terminer, signalons qu'un autre fait marquant de l'année écoulée est la parution d'un manuel de base en Statistique Fonctionnelle (voir [2]) qui est directement issu de la première édition de la conférence IWFOS2008 (qui avait eu lieu à Toulouse) et auquel ont contribué les plus grands spécialistes mondiaux par le biais de différents chapitres d'ordre méthodologique et bibliographique traitant des divers aspects de la Statistique Fonctionnelle.

[1] Ferraty, F. (Ed). Recent advances in functional data analysis and related topics, Springer, Contributions to Statistics, 2011.

[2] Ferraty, F. and Romain, Y. (Eds). The Oxford handbook on functional data analysis, Oxford University Press, 2011.

Estimation de régularité locale

Rémi Servien *

* UMR 729 MISTEA - Campus SupAgro
2 place Pierre Viala
34060 Montpellier Cedex 2.
e-mail: remi.servien@supagro.inra.fr

Résumé

Nous exposerons lors de ce groupe de travail les travaux réalisés en thèse et disponible à l'adresse suivante http://remiservien.wifeo.com/documents/thesefinale.pdf .

Le sujet principal de cet exposé est lié au problème général de dérivation des mesures (Rudin [5], Dudley [4]). Il trouve ses motivations dans l'étude de problèmes d'estimation quand les conditions de régularité habituelles ne sont pas vérifiées. En effet, de nombreux théorèmes de convergence font intervenir des hypothèses de continuité qui ne sont en pratique pas toujours satisfaites. Nous utilisons donc des conditions moins contraignantes permettant de plus d'étudier la régularité de la densité associée à la mesure considérée.

Un paramètre α_x appelé *indice de régularité* apparaît lorsqu'on essaie d'étudier localement le comportement d'une fonction de densité dérivée d'une mesure quelconque. Ce paramètre de régularité étant fortement local, son estimation est difficile. Nous nous attacherons à étudier certains problèmes d'estimation non paramétrique où cet indice intervient et à trouver différents estimateurs convergents de α_x.

Nous considérons μ une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Notons λ la mesure de Lebesgue sur $\mathbb{R}^d$ muni d'une norme notée $||.||$. Soit x un point de $\mathbb{R}^d$, δ un réel positif et $B_\delta(x)$ la boule ouverte de centre x et de rayon δ. Afin de mesurer le comportement local de $\mu(B_\delta(x))$ par rapport à $\lambda(B_\delta(x))$ nous pouvons considérer le quotient de ces deux mesures. Ainsi, si pour x fixé la limite suivante

$$f(x) = \lim_{\delta \to 0} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} \tag{0.1}$$

existe, alors x est appelé *point de Lebesgue* de la mesure μ. Si μ est absolument continue par rapport à λ nous pouvons sélectionner parmi toutes les densités obtenues à partir de μ, une densité particulière f, qui satisfait (1) en tout point où cette limite existe. Il est important de noter que la notion de point de Lebesgue est plus large que la notion de continuité. Elle permet donc d'élargir certains résultats en diminuant les contraintes sur les fonctions à estimer. Dans ce contexte, Berlinet et Levallois [3] définissent un point ρ-régulier de la mesure μ comme un point de Lebesgue x de μ tel que

$$\left|\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x)\right| \leq \rho(\delta), \tag{0.2}$$

où ρ est une fonction mesurable telle que $\lim_{\delta\downarrow 0}\rho(\delta) = 0$.

Par exemple, si $d = 1$ et si la mesure μ a une densité f avec une dérivée f' bornée par une constante quelconque C_x dans le voisinage de x, alors nous avons $\rho(\delta) = C_x\delta$ et x est ρ-régulier. Il est aussi clair que, si f est une fonction localement hölderienne en x avec un exposant α_x, cela implique $\rho(\delta) = C_x/(\alpha_x + 1)\delta^{\alpha_x}$. De plus, il est possible de trouver des exemples de mesures ρ-régulières mais avec un mauvais comportement local de la densité, comme des discontinuités du second ordre. Pour des exemples, nous renvoyons le lecteur à l'article de Berlinet et Levallois [3].

Précisons que la fonction ρ de (2) n'est pas unique et dépend de la norme choisie sur $\mathbb{R}^d$. Il est par ailleurs possible d'aller plus loin que la relation (2) et de considérer qu'en x, point de Lebesgue de la mesure μ, nous ayons

$$\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ quand } \delta \downarrow 0, \tag{0.3}$$

où C_x est une constante différente de 0 et α_x un nombre réel strictement positif que nous appellerons *indice de régularité*. Ces constantes sont alors uniques et, trivialement, cette relation implique la ρ-régularité en x avec $\rho(\delta) \sim C_x\delta^{\alpha_x}$. Cette relation définissant l'indice de régularité joue un rôle central tout au long de cet exposé. Notons que l'indice α_x reflète le degré de régularité de la mesure μ par rapport à la mesure de Lebesgue λ. En effet, plus α_x sera grand, plus la dérivée de μ sera lisse autour du point x.

La connaissance de cet indice est intéressante en pratique pour étudier le comportement local de la mesure. En effet, il nous donne d'importantes indications sur le caractère plus ou moins lisse d'une mesure autour du point x. Il est important de noter que α_x existe dans le cas d'une densité non nécessairement continue, ceci nous garantissant un large cadre de travail. Il intervient également dans différents problèmes d'estimation intimement liés au caractère lisse ou non lisse de la mesure. Afin d'étudier certains de ces problèmes, nous structurons notre présentation en trois parties.

1 Normalité asymptotique d'estimateurs de la densité

En notant $B_n(x) = B(x, R_n(x))$ la plus petite boule fermée de centre x contenant au moins k_n points de l'échantillon, Berlinet et Levallois [3] démontrent le résultat suivant : sous les conditions de convergence de f_{k_n}

$$\lim_{n\to\infty} k_n = \infty \quad \text{et} \quad \lim_{n\to\infty} \frac{k_n}{n} = \infty,$$

si x est un point ρ-régulier de la mesure μ avec $f(x) > 0$, alors la condition

$$\sqrt{k_n}\rho(R_n(x)) \xrightarrow{P} 0$$

lorsque n tend vers l'infini implique la convergence en distribution de la variable aléatoire

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

vers une loi $\mathcal{N}(0, 1)$.

Ils obtiennent ensuite comme corollaire de ce résultat que sous les conditions de convergence évoquées précédemment et si de plus la densité f est lipschitzienne d'ordre $\alpha > 0$ alors la condition

$$\lim_{n\to\infty} \frac{k_n^{1+1/2\alpha}}{n} = 0$$

implique la convergence en distribution de $T_n(x)$ vers une loi $\mathcal{N}(0, 1)$.
Une bonne estimation de l'indice de régularité est donc primordiale afin d'obtenir la vérification des conditions et, par conséquent, la normalité asymptotique de $T_n(x)$. En effet, comme nous pourrons le remarquer sur un exemple, en ne tenant pas compte de la spécificité de la mesure étudiée et de son indice de régularité il est possible de commettre une importante erreur d'estimation de la densité. Nous nous attachons tout d'abord dans ce chapitre à élargir les résultats de Berlinet et Levallois [5] sur l'estimateur des k_n-plus proches voisins. Nous obtenons une condition nécessaire et suffisante à la normalité asymptotique de la loi limite de $T_n(x)$ ainsi que l'expression de cette loi. Enfin, nous testons nos résultats sur des données simulées.

2 Estimation du mode pour des densités non continues

Considérons une mesure de probabilité μ dans $\mathbb{R}^d$ à partir de laquelle nous obtenons une densité f. Nous nous intéressons au problème de l'estimation du mode θ de f à partir d'un

échantillon i.i.d. $S_n = \{X_1, \ldots, X_n\}$ distribué selon μ. Ce problème a suscité une littérature considérable comme nous pourrons le voir plus en détails dans l'introduction de cette seconde partie. Néanmoins, tous les estimateurs nécessitent des conditions de régularité forte, au minimum la continuité autour du mode θ. Ainsi, pour tout $x \in \mathbb{R}^d$, on définit l'estimateur à noyau de la densité f_{h_n} par

$$f_{h_n}(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_n}\right),$$

où k est un noyau et h_n la fenêtre de lissage strictement supérieure à 0 et telle que h_n tend vers 0 quand n tend vers l'infini. Abraham, Biau et Cadre [1] définissent un estimateur θ_n du mode par

$$\theta_n \in \arg\max_{S_n}(f_{h_n}),$$

plus précisément,

$$\theta_n \in \left\{ x \in S_n : f_{h_n}(x) = \max_{1 \leq i \leq n} f_{h_n}(X_i) \right\}.$$

Cet estimateur converge alors presque sûrement vers le mode θ si f est continue autour du mode. Ils obtiennent ensuite une bonne vitesse de convergence et un intervalle de confiance asymptotique. En nous appuyant sur leurs travaux, nous démontrons la convergence de θ_n sous des conditions ne portant pas sur la régularité de la densité f. Nous montrons ensuite que ces conditions sont vérifiées pour $\theta \in V$ où V est un intervalle où tout point est un point ρ-régulier. Nous obtenons également des intervalles de confiance asymptotiques sous certaines hypothèses supplémentaires, qui sont vérifiées pour θ appartenant à un intervalle de points de Lebesgue admettant un indice de régularité.

3 Estimateurs de l'indice de régularité utilisant des estimateurs de la fonction de répartition

Nous avons pu remarquer que l'indice de régularité intervient dans différents problèmes d'estimation non paramétrique. Cependant, il est difficile à estimer et le seul estimateur disponible à notre connaissance est celui de Beirlant, Berlinet et Biau [2] qui utilise l'estimateur des k_n-plus proches voisins de la densité. Ils définissent leur estimateur $\bar{\alpha}_{n,x}$, quelque soit $\tau > 1$, par

$$\bar{\alpha}_{n,x} = \frac{d}{\log \tau} \log \frac{f_{\lfloor \tau^2 k_n \rfloor}(x) - f_{\lfloor \tau k_n \rfloor}(x)}{f_{\lfloor \tau k_n \rfloor}(x) - f_{\lfloor k_n \rfloor}(x)},$$

si $[f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)]/[f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)] > 1$ et $\bar{\alpha}_{n,x} = 0$ sinon, $\lfloor . \rfloor$ étant la fonction partie entière. Ils obtiennent la convergence en probabilité et la normalité asymptotique de cet estimateur. Cependant, les simulations s'avèrent perfectibles. En nous appuyant sur cet article, nous déterminons dans la troisième partie un nouvel estimateur de l'indice de régularité en utilisant des estimateurs de la fonction de répartition. Ainsi, nous obtenons un nouvel estimateur convergent, sous certaines hypothèses, de l'indice de régularité

$$\alpha_{n,x} = \frac{d}{\log \tau} \log \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)},$$

où

$$\varphi_{n,\delta_n}(x) = \frac{\mu_n(B_{\delta_n}(x))}{\lambda(B_{\delta_n}(x))}$$

avec μ_n la mesure empirique. Enfin, nous terminons ce chapitre par une étude pratique sur des données simulées.

Références

[1] C. Abraham, G. Biau et B. Cadre. Simple estimation of the mode of a multivariate density. *The Canadian Journal of Statistics*, 23-34, 2003.

[2] J. Beirlant, A. Berlinet and G. and Biau. Higher order estimation at Lebesgue points. *Annals of the Institute of Statistical Mathematics*, 60:651-677, 2008.

[3] A. Berlinet and S. Levallois. Higher order analysis at Lebesgue points. In M.L. Puri, editor, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, 1-16, 2000.

[4] R. Dudley. *Real Analysis and Probability*, Chapman and Hall, New-York, 1989.

[5] W. Rudin. *Real and Complex Analysis*, McGraw-Hill, New-York, 1987.

.

Structure de la mesure aléatoire associée à un processus isotrope

Alain Boudou* et Sylvie Viguier-Pla*

* Université Paul Sabatier
Institut de Mathématiques de Toulouse
UMR 5219
118 route de Narbonne
F-31062 Toulouse Cedex 9.
e-mail: boudou@math.univ-toulouse.fr, viguier@math.univ-toulouse.fr

Résumé

A tout processus stationnaire on peut associer, d'une façon biunivoque, une mesure aléatoire dont il est la transformée de Fourier. Ainsi, toute particularité d'un processus dans le domaine temporel a sa traduction dans le domaine fréquentiel.

Ici, nous nous proposons d'étudier la forme de la mesure aléatoire lorsque le processus est isotrope. Nous sommes ainsi amenés à définir le produit tensoriel de mesures aléatoires.

Références

Boudou, A. (2007). Groupe d'opérateurs unitaires déduit d'une mesure spectrale - une application. *C. R. Acad. Sci. Paris, Ser. I* **344** 791-794

Boudou, A. and Romain, Y. (2002). On spectral and random measures associated to continuous and discrete time processes. *Stat. Proba. Letters* **59** 145-157.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory.* 2nd ed. Society for Industrial Applied Mathematics, Philadelphia.

Dacunha-Castelle, D. and Duflo, M. (1982). *Probabilités et Statistiques.* Masson.

Dehay, D. and Monsan, V. (2007) Discrete Periodic Sampling with Jitter and Almost Periodically Correlated Processes. *Stat. Infer. Stoch. Process.* **10** 223-253.

Matilla, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces.* Cambridge, University press.

Riesz, F. and Nagy, B. (1991). *Functional Analysis.* Dover Publications.

Rozanov, Yu. A. (1967). *Stationary Random Processes.* Holden-Day, Inc, San Francisco.

Shumway, R. H. and Stoffer, Da. S. (2006). *Stationary Time Series Analysis and its Applications.* Springer, New-York.

Estimation of the Semiparametric Factor Model: Application to Modelling Time Series of Electricity Spot Prices.

Dominik Liebl *

* Adresse pour correspondance:
Seminar für Wirtschafts– und Sozialstatistik,
Lehrstuhl Prof. Mosler, Universität zu Köln, Albertus–Magnus–Platz, 50923 Köln Germany
liebl@statistik.uni-koeln.de

Introduction

Classical univariate and multivariate time series models have problems to deal with the high variability of hourly electricity spot prices. We propose to model alternatively the daily mean electricity supply functions using a dynamic factor model. And to derive, subsequently, the hourly electricity spot prices by the evaluation of the estimated supply functions at the corresponding hourly values of demand for electricity. Supply functions are price (EUR/MWh) functions, that increase monotonically with demand for electricity (MW). Apart from this new conceptual approach, that allows us to represent the auction design of energy exchanges in a most natural way, our main contribution is an extraordinary simple algorithm to estimate the factor structure of the dynamic factor model. We decompose the time series into a functional "spherical component" and an univariate "scaling component". The elements of the spherical component are all standardized having unit size such that we can robustly estimate the factor structure. This algorithm is much simpler than procedures suggested in the literature. In order to use a parsimonious labeling we will refer to the daily mean supply curves simply as "price curves".

The Dynamic Semiparametric Factor Model (DSFM) of [7] and the follow up application to electricity spot prices in [5] are close to our approach, but there are two important differences. Firstly, the authors model the hourly spot prices directly as a multivariate time series and therein fail to mirror the auction design (i.e. the data generating process) at electricity exchanges. As a result, they are able to explain only about 80% of the variation in hourly spot prices at the European Electricity Exchange while we are able to explain over 98% of the variation using the same number of factors. Secondly, they use an iterating optimization

algorithm to estimate the factor structure, whereas we use principal component analysis for sparse functional data [2] to estimate the factor structure of the spherical component. And we show that the estimated factor structure of the spherical component is also the factor structure of the original series.

Functional Dynamic Factor Model

We model the prices, Y_{ti}, as observations of an underlying smooth price curve, X_t, such that

$$Y_{ti} \quad = \quad X_t(u_{ti}) + \varepsilon_{ti} \quad \text{with } t = 1, 2, \ldots, T. \tag{3.1}$$

Where $X_t(.)$ is a smooth monotone random function of adjusted demand[1] $u \in \mathcal{U}$ with $\mathcal{U}$ being a closed and bounded subspace of $\mathbb{R}$. We will set, without loss of generality, $\mathcal{U} = [0, 1]$. The index $i = 1, \ldots, N_t$ in u_{ti} refers to the i-th order statistic of the observed hourly adjusted demand values, u_{th}. The noise term, ε_{ti}, is assumed to be independently distributed for each t and i, with $\mathrm{E}(\varepsilon_{ti}) = 0$ and $\mathrm{Var}(\varepsilon_{ti}) = \sigma_\varepsilon^2$. An example of some raw data vectors $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{tN_t})'$ can be seen in figure 1. Note that, some prices Y_{ti} have to be treated as outliers, and we use N_t to refer to the amount of prices per day t, that is used in the estimation procedure. An example of outlier prices can be seen in the left panel of figure 3.

Dynamic factor models are a very successful approach to analyze high dimensional time series data. Our case is a special case of the generalized dynamic factor models considered in [8] and corresponds to the dynamic factor model in [7]. The factor structure, F, consists of unknown non parametric functions, $f_1, \ldots, f_K$, that have to be estimated from the data. The $K < \infty$ functionals of the estimated factor structure, $\hat{F} = [\hat{f}_1, \ldots, \hat{f}_K]$, are required to be mutually orthonormal to each other and to be an optimal empirical basis such that

$$X_t \approx \sum_{k=1}^{K} \hat{\beta}_{tk} \hat{f}_k = \hat{\beta}_t' \hat{F}. \tag{3.2}$$

More precisely, the factor structure, $\hat{F} = [\hat{f}_1, \ldots, \hat{f}_K]$, shall define the best possible projection from the space $\mathcal{H}_T \subset L^2(\mathcal{U})$ spanned by the sampled functions, $X_1, \ldots, X_T$, into a K dimensional subspace of $\mathcal{H}_T$, where "best possible" is understood with respect to the mean squared error sense,

$$\sum_{t=1}^{T} ||X_t - \sum_{k=1}^{K} \hat{\beta}_{tk} \hat{f}_k||_2^2 = \min_{v_1, \ldots, v_K} \sum_{t=1}^{T} \min_{\vartheta_1, \ldots, \vartheta_K} ||X_t - \sum_{k=1}^{K} v_{tk} \vartheta_k||_2^2, \tag{3.3}$$

[1] Adjusted demand means: Original demand values minus electricity from wind-power. Because of its privileged status of renewable energy sources, the market price of electricity is not valid for wind-power.

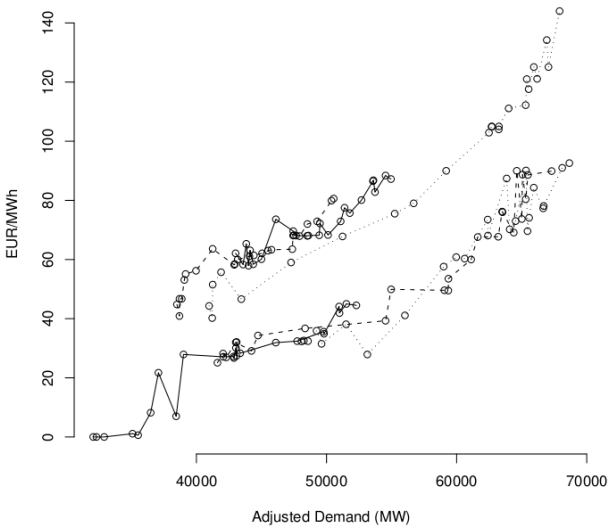

Figure 1: Three consecutive days from two different arbitrary weeks.

with respect to all possible $\vartheta_1, \ldots, \vartheta_t \in L^2(\mathcal{U})$ and $v_{t1}, \ldots, v_{tK} \in \mathbb{R}$. We use $||.||_2$ to denote the L2-norm, in its functional version $||f||_2 = \sqrt{\int_0^1 f(u)^2 du}$ for functions $f \in L^2(\mathcal{U})$, and its euclidean version $||y||_2 = \sqrt{\sum_{i=1}^N y_i^2}$ for vectors $y \in \mathbb{R}^N$. Note that this definition of a factor structure, $\hat{F}$, is also fulfilled by any rigid rotations, $\hat{F}^* = \mathbf{T}\hat{F}$, where $\mathbf{T}$ is any orthonormal $K \times K$-matrix such that $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = I_K$.

It is well known that the first $K < \infty$ empirical eigenfunctions, let's say $f_{1T}, \ldots, f_{KT}$, of the sample covariance operator,

$$\begin{aligned} \rho_T\, g &= \int_0^1 \sigma_T(u,v)\, g(v) dv, \quad \text{for all } \; g \in L^2(\mathcal{U}), \\ \text{where} \quad \sigma_T(u,v) &= T^{-1} \sum_{t=1}^{T} X_t(u) X_t(v), \quad \text{with } \; u, v \in \mathcal{U}, \end{aligned}$$

can define such a best possible projection from the space $\mathcal{H}_T = \text{span}(X_1, \ldots, X_T) \subset L^2(\mathcal{U})$ into a K-dimensional subspace of $\mathcal{H}_T$. In our general setting, where $(X_1, \ldots, X_T)$ is allowed to be any collection of functional random variables the sample covariance operator, $\rho_T\, g$, generally does not converge to a population counterpart and the empirical eigenfunctions and eigenvalues cannot be interpreted as variance components in the classical sense. This

sample dependence of $F_T = [f_{1T}, \ldots, f_{KT}]$ is not different from other dynamic factor models as in [7].

Unfortunately, given the unrestrictive assumptions on the series (X_t), the spectral decomposition of the empirical covariance operator, $\rho_T\, g$, generally cannot be used to estimate a factor structure, F_T. As long as the process (X_t) is not stationary, its elements are likely to be of very different orders of magnitude, which will have a dramatic distortion effect on the sample covariance function, σ_T. But, contrary to the claim of the authors in [7], we do not need stationarity in order to use spectral decomposition of the sample covariance operator to estimate a factor structure for the functions $X_1, \ldots, X_T$.

Proposition 1 *Given the model in (3.2), if a factor structure $\hat{F}$ defines the best projection from the space $\mathcal{H}_T = span(X_1, \ldots, X_T)$ into a K dimensional subspace $\mathcal{H}_T^K \subset \mathcal{H}_T$, then it also defines the best projection from the space $\mathcal{H}_T^* = span(\frac{X_1}{||X_1||_2}, \ldots, \frac{X_T}{||X_T||_2})$ into the same K dimensional subspace $\mathcal{H}_T^K$.*

This proposition is trivially true, because $\mathcal{H}_T = \text{span}(X_1, \ldots, X_T)$ is a vector space and therefore is closed under scalar multiplication, such that $\mathcal{H}_T = \mathcal{H}_T^*$. Different scales $X_t\, c_t$, with $c_t \neq 0$, will simply cause reciprocal scales of $\hat{\beta}/c_t$ in the minimization (3.3). As a consequence from proposition 1 we can also estimate a factor structure for the original series, (X_t), from the standardized series $(\frac{X_t}{||X_t||_2})$.

The Algorithm

The idea is to decompose the time series, $(\mathbf{Y}_t)$, into its "spherical" component that can be used to estimate the K-dimensional factor structure F and its "scaling" component that can be used to rescale the approximated spherical process to its original size.

Definition 1 *The spherical component of the factor model in equation (3.2) is given by the multivariate series,*

$$\left(\frac{\mathbf{Y}_t - \mu_T(u_t)}{||\mathbf{Y}_t - \mu_T(u_t)||_2} \right)_{t=1,\ldots,T}. \tag{3.4}$$

With $u_t = (u_{t1}, \ldots, u_{N_t 1})$ and $\mu_T = T^{-1} \sum_{t=1}^{T} X_t$ being the sample mean function.

Definition 2 *The scaling component is given by the univariate series,*

$$(||\mathbf{Y}_t - \mu_T(u_t)||_2)_{t=1,\ldots,T}\,. \tag{3.5}$$

>From a mathematical perspective, it is not necessary to subtract the sample mean, $\mu_T \in \mathcal{H}_T = \text{span}(X_1, \ldots, X_T)$, from the discretization vectors, $\mathbf{Y}_T$. This simply subtracts the constant vector $\bar{\hat{\beta}} = (T^{-1}\sum_t^T \hat{\beta}_{t1}, \ldots, T^{-1}\sum_t^T \hat{\beta}_{tK})'$ from the process $(\hat{\beta}_t) = (\hat{\beta}_{t1}, \ldots \hat{\beta}_{tK})'$. But, from a practical perspective, the subtraction of the sample mean, μ_T, helps to avoid rounding errors caused by floating point computation. Particularly, when the sizes of different vectors $\mathbf{Y}_t$ are of very different orders of magnitude, as in our application.

By construction, the elements of the spherical component, $\left(\frac{\mathbf{Y}_t - \mu_T(u_t)}{||\mathbf{Y}_t - \mu_T(u_t)||_2}\right)$, are all of the same order of magnitude, such that the factor structure, F, can be estimated by the spectral decomposition of the spherical sample covariance operator,

$$\begin{aligned} \tilde{\rho}_T\, g &= \int_0^1 \tilde{\sigma}_T(u,v)\, g(v) dv, \quad \text{for all } \; g \in L^2(\mathcal{U}), \\ \text{where} \quad \tilde{\sigma}_T(u,v) &= T^{-1}\sum_{t=1}^T \frac{\mathbf{Y}_t(u) - \mu_T(u)}{||\mathbf{Y}_t(u) - \mu_T(u)||_2} \frac{\mathbf{Y}_t(v) - \mu_T(v)}{||\mathbf{Y}_t(v) - \mu_T(v)||_2}, \end{aligned}$$

without distortion effects. This estimation algorithm is by far less costly with respect to computation time and much simpler to implement than the iterative procedure in [7].

Application

The estimation of a factor structure, F, for the daily mean electricity supply functions, X_t, is made a bit more difficult by the sparseness of the data. The observed discretization points, $\mathbf{Y}_t$, of the price functions, X_t, are not uniformly distributed over the whole domain $\mathcal{U} = [0,1]$, but over sub parts of $\mathcal{U}$. This is a slightly different form of sparseness as it is discussed in [4] and [2], where sparseness is referred to the situation with only a few discretization points per function. Nevertheless the smoothing approaches suggested by [4], to estimate the mean function and the covariance operator, as well as the PACE estimation procedure of [2], to estimate the loadings parameters, are directly applicable to our situation of sparse data. The empirical covariance function, $\tilde{\sigma}_T$, and the first four factors, $f_{1T}, \ldots, f_{4T}$, can be seen in figure 2. The estimated factor structure explains about 98.5% of the total variance of the price curves, such that we can reduce the high dimensional problem to a $K = 3$-dimensional problem without much loss of generality.

In the left panel of figure 3 we plot one estimated price function, $\hat{X}_t$, of an arbitrary day, t, with its corresponding raw data vector, $\mathbf{Y}_t$, as well as two outlier prices, that are excluded from the estimation procedure. In the right panel of figure 3 we show hourly electricity spot prices of one arbitrary week. The hourly fitted prices are determined by the evaluation of the estimated price functions, $\hat{X}_t$, at the corresponding hourly values of adjusted demand, u_{th}, for electricity. Note, that the proposed dynamic factor model may be easily combined

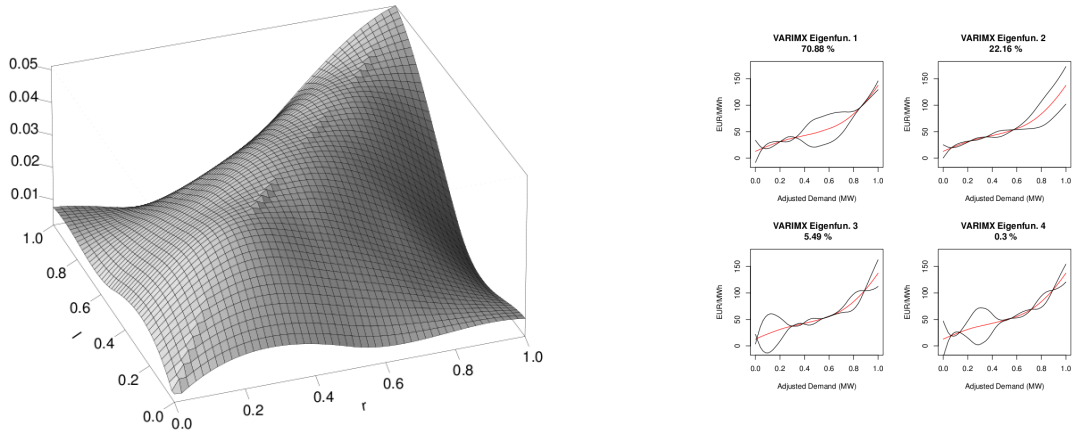

Figure 2: LEFT PANEL Empirical covariance function, $\tilde{\sigma}_T$, of the spherical component. RIGHT PANEL First four functionals of the estimated factor structure.

with already developed approaches to model and forecast demand for electricity such as in [3].

References

[1] Locantore, N and Marron, J S and Simpson, D G and Tripoli, N and Zhang, J T and Cohen, K L and Boente, G and Fraiman, R and Brumback, B and Croux, C. (2009). Robust principal component analysis for functional data, Test, 8, 1-73.

[2] Yao, F and Müller, H G and Wang, J L. (2005). Functional Data Analysis for Sparse Longitudinal Data. Journal of the American Statistical Association, 470, 577-590.

[3] Antoch, J and Prchal, L and Rosa, MR and Sarda, P. (2008). Functional linear regression with functional response: Application to prediction of electricity consumption. In *Functional and Operatorial Statistics*, 5219-5219.

[4] Staniswalis, J G and Lee, J J. (1998). Nonparametric Regression Analysis of Longitudinal Data. Journal of the American Statistical Association, 444, 1403-1418.

[5] Härdle, W K and Trück, S. (2010). The dynamics of hourly electricity prices. Preprint.

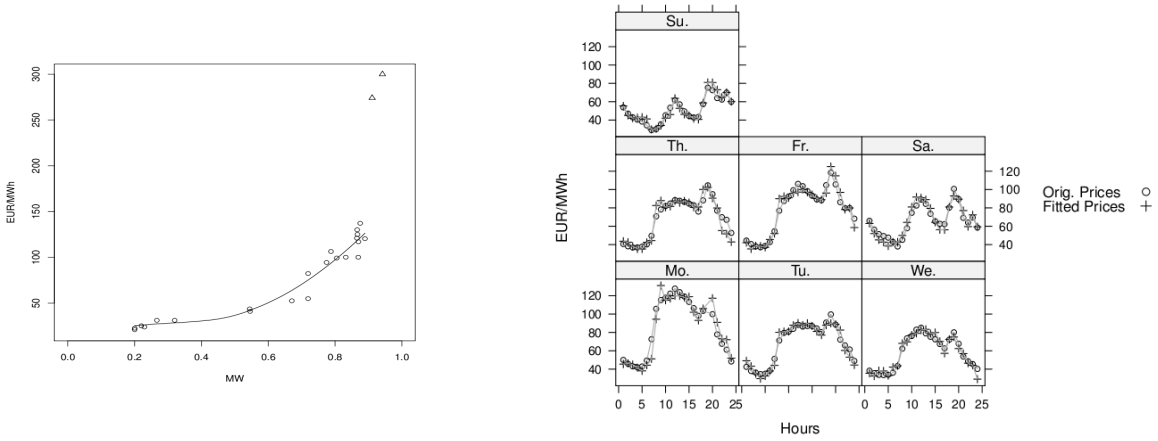

Figure 3: Left Panel Single fitted price curve with observed raw prices (circle points) and outlier prices (triangle points). Right Panel Hourly fitted prices and original prices.

[6] Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. Biometrika, 95, 587-600.

[7] (2009) Park, B U and Mammen, E and Härdle, W and Borak, S. (2009). Time Series Modelling With Semiparametric Factor Dynamics. Journal of the American Statistical Association, 485, 284-298.

[8] Forni, M and Hallin, M and Lippi, M and Reichlin, M. (2000). The generalized dynamic factor model: Identification and estimation. The Review of Economics and Statistics, 82, 540-554.

.

Sur les modèles non paramétriques conditionnels en statistique fonctionnelle

Amel Tadj

* Université Paul Sabatier et Institut de Mathématiques de Toulouse
UMR 5219, 118 route de Narbonne, F-31062 Toulouse Cedex 9.
e-mail: ameltdz@yahoo.fr

Cet exposé consiste en la soutenance de thèse devant le jury composé de:

Mohamed Attouch Univ. Sidi Bel Abbès Examinateur

Hervé Cardot Univ. Bourgogne Rapporteur

Frédéric Ferraty Univ. Toulouse Directeur de Thèse

Ali Gannoun Univ. Montpellier Examinateur

Stéphane Girard Univ. Grenoble Rapporteur

Aldo Goia Univ. Piemonte Orientale Examinateur

Ali Laksaci Univ. Sidi Bel Abbès Directeur de Thèse

Philippe Vieu Univ. Toulouse Directeur de Thèse

Résumé

La problématique abordée dans cette thèse est l'estimation non paramétrique des modèles conditionnels à variable explicative fonctionnelle en traitant deux cas: le cas où la variable réponse est réelle et le cas d'une variable réponse fonctionnelle. On établit la convergence uniforme presque complète d'estimateurs non paramétriques pour certains modèles conditionnels.

Dans un premier temps, nous considérons une suite d'observations i.i.d. et nous construisons des estimateurs par la méthode du noyau pour la fonction de régression généralisée, la fonction de répartition conditionnelle, la densité conditionnelle, la fonction de hasard conditionnelle et le mode conditionnel. Nous étudions la convergence uniforme presque complète de ces estimateurs en précisant leurs vitesses. A titre illustratif, nous donnons des exemples d'applications sur des données simulées.

Dans un second temps, on généralise nos résultats au cas d'une variable réponse fonctionnelle (appartenant à un espace de Banach) et on estime la régression classique. Cette généralisation a été étudiée dans les deux cas: les observations i.i.d. ainsi que le cas dépendant. Dans ce dernier, nous avons fixé comme objectif la convergence presque complète ponctuelle lorsque les observations sont β-mélangeantes.

Nos résultats asymptotiques exploitent bien la structure topologique de l'espace fonctionnel de nos observations et le caractère fonctionnel de nos modèles. En effet, toutes nos vitesses de convergence sont quantifiées en fonction de la concentration de la mesure de probabilité de la variable fonctionnelle, de l'entropie de Kolmogorov et du degré de régularité des modèles. Notons également que dans le cas où la variable réponse est aussi fonctionnelle, nos vitesses de convergence contiennent un terme additionnel qui dépend du type de l'espace de Banach de la variable réponse.

Summary

In this thesis, we consider the problem of the nonparametric estimation in the conditional models when the regressor takes its values in infinite dimension space. More precisely, we treated two cases when the response variable is real and functional. One establishes almost complete uniform convergence of nonparametric estimators for certain conditional models.

Firstly, we consider a sequence of i.i.d. observations. In this context, we build kernel estimators of the conditional cumulative distribution, the conditional density, the conditional hazard function and the conditional mode. We give the uniform consistency rate of these estimators. We illustrate our results by giving an application on simulated samples.

Secondly, we generalize our results when the response variable is in a Banach space. We estimate the regression function. In this context, we treat both cases: i.i.d and dependent observations. In the last case, we consider that the observations are β-mixing and we establishes almost complete pointwise convergence.

Our asymptotic results exploit the topological structure of functional space for the observations. Let us note that all the rates of convergence are based on an hypothesis of concentration of the measure of probability of the functional variable on the small balls and

also on the Kolmogorov's entropy which measures the number of the balls necessary to cover some set. Moreover, when the response variable is functional the rate of convergence contains a new term which depends on type of Banach space.

Liste des travaux

F. Ferraty, A. Laksaci, A. Tadj, P. Vieu. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *J. Statist. Plann. Inference.* 140, 335-352.

F. Ferraty, A. Laksaci, A. Tadj, P. Vieu. (2011). Kernel regression with functional response. *Electronic Journal of Statistics.* 5, 159-171.

F. Ferraty, A. Laksaci, A. Tadj, P. Vieu. Estimation de la régression pour variable explicative et réponse fonctionnelles dépendantes (Soumis pour publication).

Communications

F. Ferraty, A. Laksaci, A. Tadj, P. Vieu. Rate of uniform consistency for nonparametric estimates with functional variables. 5èmes Journées de Statistique Fonctionnelle et Opérationnelle (STAPH), Dijon, les 18 et 19 Juin 2009.

F. Ferraty, A. Laksaci, A. Tadj, P. Vieu. Modèle de régression non paramétrique fonctionnel à variable réponse Banachique. Journées Internationales de Statistique Théorique et Appliquée, Sidi Bel Abbés, les 10, 11 et 12 Avril 2010.

.

Discrimination de courbes de densités : une application au dépistage du cancer broncho-pulmonaire

MORLAIS Fabrice

* Adresse pour correspondance:
ERI3 INSERM "Cancers & Populations"
EA 3936 Université Caen
Unité de Recherche et d'Evaluation en Epidémiologie
Pôle de Santé des Populations
Faculté de médecine - avenue côte de nacre - 14032 Caen cedex
e-mail: fabrice.morlais@unicaen.fr

Résumé

Le dépistage du cancer broncho-pulmonaire chez des personnes ayant été exposées professionnellement à l'amiante est généralement réalisé à l'aide d'une radiographie pulmonaire, d'un scanner thoracique et d'un examen cytologique des expectorations. 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. De récentes études (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 de la lame, un certain nombre de paramètres de forme, de texture et d'intensité sont mesurés.

Pour discriminer les personnes ayant un cancer broncho-pulmonaire des personnes saines de cancer broncho-pulmonaire des personnes, nous avons développé un modèle de discrimination fonctionnelle non paramétrique comparant les distributions cellulaires (densité de probabilité).

Références

1. 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.

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

3. Ferraty, F. and Romain. Handbook on functional data analysis and related topics. *Oxford University Press*. 2011.

4. Ferraty F, Vieu P. Non Parametric Functional Data Analysis Theory and Practice. *Springer*. 2006.

5. 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.

6. 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.

7. Ramsay JO, Silverman BW. Applied Functional Data Analysis Methods and Case Studies. *Springer*. 2002.

8. Ramsay JO, Silverman BW. Functional Data Analysis Second Edition. *Springer*. 2005.

9. Silverman BW. Density Estimation for Statistics and Data Analysis. *Chapman & Hall/CRC* 1986.

10. Wasserman L. All of Nonparametric Statistics. *Springer*. 2007.