Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 336 (2003) 1025–1028
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Modèle à indice fonctionnel simple

Frédéric Ferraty, Agnès Peuch, Philippe Vieu

Lab. Stat. & Prob., Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex, France

Reçu le 2 novembre 2002 ; accepté après révision le 7 mai 2003

Présenté par Paul Deheuvels

Résumé

Ce travail propose une extension du modèle à indice simple lorsqu'on considère une variable aléatoire explicative à valeurs dans un espace de dimension infinie. On désignera génériquement un tel modèle par modèle à indice fonctionnel simple. L'apport principal de cette note réside dans une double généralisation du modèle à indice simple. D'une part, on se place dans un cadre de v.a. fonctionnelles et d'autre part, on introduit des hypothèses sur la loi de la v.a. explicative moins restrictives que celles utilisées habituellement dans le cadre vectoriel. Des premiers résultats de convergence ponctuelle sont établis. ***Pour citer cet article : F. Ferraty et al., C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

Abstract

Single Functional Index Model. This paper concerns a generalization of the "Single Index Model" when the explanatory variable is valued in an infinite dimensional space. Such model will be called the "Single Functional Index Model". The main contribution of this study is to propose a functional framework which includes the classical vectorial case. Pointwise asymptotic properties are stated under less restrictive conditions on the law of the explanatory variable than what it is usually assumed in the vectorial case. ***To cite this article: F. Ferraty et al., C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

1. Introduction

Le modèle à indice simple est très populaire dans la communauté des économètres car il répond à deux préoccupations importantes. La première concerne la réduction de dimension puisque ce type de modèle permet d'apporter une solution au problème du fléau de la dimension. La seconde est liée à l'interprétabilité de l'indice (paramètre) introduit dans ces modèles. L'étude statistique de ces modèles, dans le cadre de v.a. explicatives vectorielles, a été amorcée par Härdle et al. [6] et plus récemment, Hristache et al. [7] fournissent à la fois de nouveaux éléments théoriques et bibliographiques. Parallèlement, se sont développés depuis peu des modèles de régression (voir Dabo-Niang [2] et Goia [5], pour des aspects théoriques et bibliographiques) dont les v.a. explicatives sont à valeurs dans des espaces de dimension infinie (appelées génériquement v.a. fonctionnelles). L'apport principal de cette Note se situe à l'intersection du modèle à indice simple et des modèles de régression

Adresse e-mail : ferraty@cict.fr (F. Ferraty).

1631-073X/03/$ – see front matter
doi:10.1016/S1631-073X(03)00239-5

pour v.a. fonctionnelles (v.a.f.). En effet, ces travaux concernent l'étude d'un modèle à indice fonctionnel simple, extension du modèle à indice simple usuel au cas où la v.a. explicative est fonctionnelle. Ce modèle peut aussi être vu comme une généralisation du modèle linéaire fonctionnel (voir Cardot et al. [1], pour des développements récents). De fait, ce modèle hérite des potentialités du modèle linéaire fonctionnel en termes d'applications (voir Ramsay et Silverman [8]) et de celles du modèle fonctionnel de régression nonparamétrique (voir Ferraty et Vieu [4]).

Le prochain paragraphe définit le modèle à indice fonctionnel simple et traite brièvement du problème d'identifiabilité. Le troisième paragraphe présente des premiers résultats de convergence ponctuelle dans un cadre général. Enfin, le dernier paragraphe est consacré aux commentaires suscités par une telle approche ainsi qu'aux problèmes ouverts.

2. Modèle à indice fonctionnel simple

L'objectif de ce paragraphe est de proposer une extension du modèle à indice simple afin de prendre en compte une v.a.f. explicative. Plus précisément, soit la v.a. Y considérée comme une réponse scalaire et X une v.a.f. explicative, supposée à valeurs dans un espace hilbertien séparable $\mathcal{H}$ muni du produit scalaire $\langle\cdot,\cdot\rangle$. On dispose de n couples $(X_i, Y_i)_{i=1,\dots,n}$ identiquement et indépendamment distribués selon la loi de (X, Y). Le modèle à indice fonctionnel simple est défini par :

$$Y_i = r\big(\langle\theta_0, X_i\rangle\big) + \varepsilon_i, \quad \forall i = 1, \dots, n,$$

où r est une fonction réelle, $\theta_0 \in \Theta \subset \mathcal{H}$ et pour $i = 1, \dots, n$, ε_i est une v.a.r. telle que $E(\varepsilon_i | X_i) = 0$.
Le premier problème que l'on doit résoudre concernant un tel modèle est celui de l'identifiabilité. Dans ce but, considérons $(e_1, \dots, e_i, \dots)$ une base orthonormée de $\mathcal{H}$. On peut alors énoncer :

Proposition 2.1. *Si r et r^* sont dérivables et si $\Theta = \{\theta \in \mathcal{H} \mid \langle\theta, e_1\rangle = 1\}$, alors*

$$\forall x \in \mathcal{H}, \quad r\big(\langle x, \theta\rangle\big) = r^*\big(\langle x, \theta^*\rangle\big) \quad \Rightarrow \quad r \equiv r^* \text{ et } \theta \equiv \theta^*.$$

Autrement dit, le modèle à indice fonctionnel simple est identifiable.

Schéma de la preuve. Posons $x = \sum_{i=1}^{+\infty} x_i e_i$, $\theta = \sum_{i=1}^{+\infty} \theta_i e_i$ et $\theta^* = \sum_{i=1}^{+\infty} \theta_i^* e_i$. On a donc $\langle x, \theta\rangle = \sum_{i=1}^{+\infty} x_i\theta_i$ et $\langle x, \theta^*\rangle = \sum_{i=1}^{+\infty} x_i\theta_i^*$. Or, en dérivant par rapport à x_i pour $i = 1, 2, \dots$, il vient que

$$\forall x \in \mathcal{H}, \quad r\big(\langle x, \theta\rangle\big) = r^*\big(\langle x, \theta^*\rangle\big) \quad \Rightarrow \quad \theta_i r'\left(\sum_{j=1}^{+\infty} x_j\theta_j\right) = \theta_i^* r^{*\prime}\left(\sum_{j=1}^{+\infty} x_j\theta_i^*\right).$$

Comme $\theta_1 = \theta_1^* = 1$, on peut montrer que $\theta \equiv \theta^*$ et, par suite, que $r \equiv r^*$.

3. Estimateur et étude asymptotique

Tout d'abord, soit $r_{\theta_0}(\cdot) = r(\langle\cdot, \theta_0\rangle)$ l'opérateur de $\mathcal{H}$ dans $\mathbb{R}$; remarquons que $r_{\theta_0}(X) = E(Y|\langle X, \theta_0\rangle)$. En s'inspirant des estimateurs à noyau de la régression, on propose un estimateur de r_{θ_0}, noté $\hat{r}_{\theta_0}$ et défini par :

$$\hat{r}_{\theta_0}(x) = \frac{\sum_{i=1}^n Y_i K\left(h^{-1}\langle X_i - x, \theta_0\rangle\right)}{\sum_{i=1}^n K\left(h^{-1}\langle X_i - x, \theta_0\rangle\right)}, \quad \forall x \in \mathcal{H},$$

où K est un noyau et $h = h_n$ une suite de nombres positifs. Cet estimateur est une généralisation au cadre fonctionnel de celui utilisé dans le cas vectoriel (voir Delecroix et Hristache [3]).

Les résultats que nous exposons ci-dessous concernent le cas où l'indice θ_0 est fixé. Cependant nous donnerons à la fin du dernier paragraphe quelques pistes pour estimer θ_0 dans le cas où il est inconnu. Avant d'aller plus loin, introduisons la v.a.r. $Z = \langle X, \theta_0 \rangle$ qui joue un rôle prépondérant dans ce type de modélisation. Or, les résultats asymptotiques dans le cas vectoriel nécessitent l'introduction de conditions de régularité sur la densité de la v.a. Z ainsi construite. Afin d'assouplir ce type d'hypothèses, plutôt que de considérer la densité de Z, nous utiliserons les accroissements de sa fonction de répartition F_Z. C'est pourquoi on pose désormais :

$$\forall x \in \mathcal{H}, \quad G_{\theta_0}(x,h) = P\big(|\langle X - x, \theta_0\rangle| < h\big) = F_Z\big(\langle x, \theta_0 > +h\rangle\big) - F_Z\big(\langle x, \theta_0\rangle - h\big).$$

Nous pouvons maintenant énoncer le résultat principal en considérant le modèle identifiable :

Théorème 3.1. *Soit K un noyau borné et strictement positif sur son support $[-1,1]$, soit Y telle que $\exists p \geqslant 2$, $E|Y|^p < \infty$, et supposons que r_{θ_0} vérifie* :

$$\exists C < \infty,\ \exists \beta > 0,\ \forall (x,y) \in \mathcal{H}^2, \quad \big|r_{\theta_0}(x) - r_{\theta_0}(y)\big| \leqslant C\|x - y\|^{\beta},$$

où $\|\cdot\|$ est la norme associée au produit scalaire $\langle\cdot,\cdot\rangle$. Alors, si h satisfait

$$\lim_{n\to\infty} h = 0 \quad \textit{et} \quad \lim_{n\to\infty} \frac{\log n}{n\, G_{\theta_0}(x,h)} = 0,$$

on a :

$$\hat{r}_{\theta_0}(x) - r_{\theta_0}(x) = \mathrm{O}\big(h^{\beta}\big) + \mathrm{O}\left(\sqrt{\frac{\log n}{n\, G_{\theta_0}(x,h)}}\right) \quad p.co.$$

Schéma de la preuve. Posons $\Delta_i(x) = K(h^{-1}\langle X_i - x, \theta_0\rangle)$ et soit :

$$\hat{r}_{\theta_0,1}(x) = \frac{1}{n\, E\Delta_1(x)} \sum_{i=1}^{n} \Delta_i(x) \quad \text{et} \quad \hat{r}_{\theta_0,2}(x) = \frac{1}{n\, E\Delta_1(x)} \sum_{i=1}^{n} Y_i \Delta_i(x).$$

On a ainsi la décomposition suivante :

$$\begin{aligned}\hat{r}_{\theta_0}(x) - r_{\theta_0}(x) \;=\; & \frac{1}{\hat{r}_{\theta_0,1}(x)}\Big\{\big(\hat{r}_{\theta_0,2}(x) - E\hat{r}_{\theta_0,2}(x)\big) - \big(r_{\theta_0}(x) - E\hat{r}_{\theta_0,2}(x)\big)\Big\} \\ & + \frac{r_{\theta_0}(x)}{\hat{r}_{\theta_0,1}(x)}\Big\{\big(\hat{r}_{\theta_0,1}(x) - E\hat{r}_{\theta_0,1}(x)\big) - \big(E\hat{r}_{\theta_0,1}(x) - 1\big)\Big\}.\end{aligned}$$

En remarquant que pour tout x dans $\mathcal{H}$, $E\hat{r}_{\theta_0,1}(x) = 1$, le résultat annoncé sera obtenu dès qu'on aura montré que $r_{\theta_0}(x) - E\hat{r}_{\theta_0,2}(x) = \mathrm{O}(h^{\beta})$, ainsi que

$$\hat{r}_{\theta_0,2}(x) - E\hat{r}_{\theta_0,2}(x) = \mathrm{O}\left(\sqrt{\frac{\log n}{n\, G_{\theta_0}(x,h)}}\right) \text{p.co.} \quad \text{et} \quad \hat{r}_{\theta_0,1}(x) - E\hat{r}_{\theta_0,1}(x) = \mathrm{O}\left(\sqrt{\frac{\log n}{n\, G_{\theta_0}(x,h)}}\right) \text{p.co.}$$

Les deux dernières formules utilisent le corollaire de Yurinskiĭ ([9], p. 491) alors que la première s'obtient en combinant la condition de régularité sur r_{θ_0} avec :

$$E\big[Y\Delta_1(x)\big] = E\big[\big(r_{\theta_0}(X_1) - r_{\theta_0}(x)\big)K\big(h^{-1}\langle X_1 - x, \theta_0\rangle\big)\big] + r_{\theta_0}(x) E\Delta_1(x).$$

4. Commentaires et problèmes ouverts

Nous commençons ce paragraphe par la description de diverses situations atypiques pour lesquelles notre étude théorique reste valide. En effet, dans le cas vectoriel, on suppose généralement que la densité de la v.a.r. $Z = \langle X, \theta_0\rangle$

est plusieurs fois continûment dérivable. Or, dans notre contexte fonctionnel, seule la fonction de répartition F_Z de la v.a.r. Z intervient, ce qui nous permet d'englober les cas suivants :

(1) F_Z admet en $z = \langle x, \theta_0 \rangle$ une discontinuité de 1ère espèce ; alors on a

$$\hat{r}_{\theta_0}(x) - r_{\theta_0}(x) = \mathrm{O}\big(h^{\beta}\big) + \mathrm{O}\Big(\sqrt{n^{-1}\log n}\Big) \quad \text{p.co.},$$

(2) Plaçons-nous maintenant dans le cas où F_Z est supposée continue en $z = \langle x, \theta_0 \rangle$ et telle que pour $\alpha > 0$, $F_Z(z+h) - F_Z(z-h) \sim C_x h^{\alpha}$. Le Théorème 3.1 nous dit alors que

$$\hat{r}_{\theta_0}(x) - r_{\theta_0}(x) = \mathrm{O}\big(h^{\beta}\big) + \mathrm{O}\Big(\sqrt{h^{-\alpha} n^{-1}\log n}\Big) \quad \text{p.co.}$$

Selon les valeurs de α, diverses situations sont ainsi prises en compte :
(a) $\alpha < 1 \Rightarrow F_Z'$ n'existe pas au point z,
(b) $\alpha = 1 \Rightarrow F_Z'(z) > 0$,
(c) $\alpha > 1 \Rightarrow F_Z'(z) = 0$.

Notons que les hypothèses classiques dans le cas vectoriel se limitent à des situations du type (2)(b).

Dans le futur, nous nous intéresserons au cas où l'indice θ_0 est inconnu. Dans ce cas, une approche consisterait à estimer θ_0 par l'élément $\hat{\theta}_0$ de Θ qui minimise le critère de validation croisée défini par :

$$\hat{\theta}_0 = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \big[Y_i - \hat{r}_{\theta}^{-i}(X_i)\big]^2,$$

où $\hat{r}_{\theta}^{-i}$ est obtenu à partir de $(X_j, Y_j)_{j\neq i}$. De plus, les aspects calculatoires peuvent être simplifiés en remplaçant θ_0 par une approximation dans un sous-espace de dimension finie de $\mathcal{H}$ (par exemple en utilisant une base de fonctions B-splines).

Remerciements

Les participants au groupe de travail « STAPH » du LSP de Toulouse sont vivement remerciés pour leurs commentaires pertinents et permanents.

Références

[1] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statist. Sinica (2003), to appear.
[2] S. Dabo-Niang, Sur l'estimation fonctionnelle en dimension infinie. Application aux diffusions, Doctorat Université Paris VI, 2002.
[3] M. Delecroix, M. Hristache, M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique, Bull. Belg. Math. Soc. Simon Stevin 6 (1999) 161–185.
[4] F. Ferraty, P. Vieu, The functional nonparametric model and application to spectrometric data, Comput. Statist. 17 (2002) 545–564.
[5] A. Goia, Modèles de régression pour des variables aléatoires fonctionnelles, Doctorat, Université Paul Sabatier, Toulouse, 2002.
[6] W. Härdle, P. Hall, H. Ichumira, Optimal smoothing in single-index models, Ann. Statist. 21 (1993) 157–178.
[7] M. Hristache, A. Juditsky, V. Spokoiny, Direct estimation of the index coefficient in the single-index model, Ann. Statist. 29 (2001) 595–623.
[8] J. Ramsay, B.W. Silverman, Functional Data Analysis, Springer-Verlag, 1997.
[9] V. Yurinskiĭ, Exponential inequalities for sums of random vectors, J. Multivariate Anal. 6 (1976) 473–499.