Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 345 (2007) 519–522
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Régression fonctionnelle sur composantes principales pour variable explicative bruitée

Christophe Crambes

Laboratoire de statistique et probabilités, Université Paul-Sabatier, 118, route de Narbonne, 31062 Toulouse cedex, France

Reçu le 15 mars 2007 ; accepté après révision le 2 octobre 2007

Disponible sur Internet le 31 octobre 2007

Présenté par Paul Deheuvels

Résumé

Cette Note a pour objet une étude du modèle linéaire fonctionnel lorsque la variable explicative est bruitée. Pour chaque courbe explicative bruitée, on utilise une méthode de lissage à noyau, puis on effectue une régression fonctionnelle sur composantes principales. On présente la procédure d'estimation du coefficient fonctionnel du modèle, ainsi qu'un résultat de convergence de l'estimateur construit. ***Pour citer cet article : C. Crambes, C. R. Acad. Sci. Paris, Ser. I 345 (2007).***

Abstract

Functional principal component regression with noisy covariate. This Note deals with a study of the functional linear model when the covariate is noisy. We smooth each noisy curve using a kernel smoothing method, and then a functional principal component regression is done. We present the estimation procedure of the functional coefficient of the model, as well as a convergence result of the estimator. ***To cite this article: C. Crambes, C. R. Acad. Sci. Paris, Ser. I 345 (2007).***

1. Introduction

Les *données fonctionnelles* ont fait leur apparition ces dernières années dans plusieurs domaines de la statistique appliquée (chimiométrie, environnement, linguistique, économie, ...), et il est de plus en plus fréquent de travailler avec des données assimilables à des courbes ou des surfaces (voir notamment [6]). Une situation particulière faisant intervenir des données fonctionnelles est celle où on cherche à expliquer une variable Y réelle à l'aide d'une variable X à valeurs dans un espace de fonctions, qui sera ici l'espace $L^2([0,1])$ des fonctions de carré intégrable sur $[0,1]$ (muni de son produit scalaire usuel $\langle\cdot,\cdot\rangle$ et de sa norme associée $\|\cdot\|$). Sans perte de généralité, on supposera que la variable X est centrée. Le modèle particulier auquel on s'intéresse ici est le *modèle linéaire fonctionnel* (voir, entre autres, [3,6]), qui s'écrit

$$Y = \langle \alpha, X\rangle + \epsilon, \tag{1}$$

Adresse e-mail : crambes@cict.fr.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.10.015

où $\alpha \in L^2([0,1])$ est le *coefficient fonctionnel* du modèle (inconnu) et ϵ est une variable aléatoire réelle d'erreur telle que $\mathbb{E}(\epsilon|X) = 0$. En supposant que l'on dispose d'un échantillon i.i.d. $(X_i, Y_i)_{i=1,\dots,n}$ issu de (X, Y), on souhaite estimer α. Cependant, il semble plus réaliste d'envisager que la variable X est entachée d'erreur, ne serait-ce que par le fait qu'elle est en pratique mesurée en certains points de l'intervalle $[0, 1]$. Dans la suite, on note $t_1 < \cdots < t_p$ ces points de mesure, que l'on supposera équirépartis sur $[0, 1]$. Ainsi, on considère que les observations disponibles pour la variable explicative sont $W_1, \dots, W_n$, telles que

$$W_i(t_j) = X_i(t_j) + \delta_{ij}, \tag{2}$$

pour tout $i = 1, \dots, n$ et tout $j = 1, \dots, p$, où $(\delta_{ij})_{i=1,\dots,n, j=1,\dots,p}$ est une suite de variables aléatoires indépendantes et identiquement distribuées représentant les erreurs faites en chaque point $t_1, \dots, t_p$. Le but est de donner une estimation de α sur la base des observations $(W_i, Y_i)_{i=1,\dots,n}$. Dans ce contexte, un estimateur de α a été proposé dans [2]. On souhaite proposer ici une autre méthode d'estimation de α.

2. Construction de l'estimateur

Dans un premier temps, on commence par lisser les courbes bruitées à l'aide d'un estimateur à noyau de type Nadaraya–Watson (voir par exemple [4]). On définit, pour $i = 1, \dots, n$ et pour $t \in [0, 1]$,

$$\widetilde{W}_i(t) = \frac{\sum_{j=1}^p W_i(t_j) K(\frac{t-t_j}{h})}{\sum_{j=1}^p K(\frac{t-t_j}{h})}, \tag{3}$$

les courbes $\widetilde{W}_1, \dots, \widetilde{W}_n$ ainsi construites donnant une estimation des courbes $X_1, \dots, X_n$ inconnues. Dans cette expression (3), le *noyau* K est une fonction définie sur $[-1, 1]$, paire et d'intégrale égale à 1 et la *largeur de fenêtre* $h > 0$ permet de contrôler le lissage de la courbe estimée $\widetilde{W}_i$.

Dans un second temps, on construit une estimation de α à l'aide d'une régression sur composantes principales fonctionnelle (présentée dans [3]). Cette méthode d'estimation est basée sur la diagonalisation de l'opérateur de covariance empirique $\Gamma_{X,n}$ défini par $\Gamma_{X,n} u = \frac{1}{n} \sum_{i=1}^n \langle X_i, u\rangle X_i$ pour tout $u \in L^2([0, 1])$, qui n'est autre que la version empirique de l'opérateur de covariance Γ_X défini par $\Gamma_X u = \mathbb{E}(\langle X, u\rangle X)$ pour tout $u \in L^2([0, 1])$. L'existence de cet opérateur est assurée sous la condition $\mathbb{E}(\|X\|^2) < +\infty$, qui est assurée par exemple lorsque $\|X\|$ est bornée p.s., ce que l'on supposera dorénavant. On note $(\lambda_r)_{r\geqslant 1}$ (resp. $(\hat{\lambda}_r)_{r\geqslant 1}$) la suite des valeurs propres de Γ_X (resp. de $\Gamma_{X,n}$) rangées dans l'ordre décroissant et $(g_r)_{r\geqslant 1}$ (resp. $(\hat{g}_r)_{r\geqslant 1}$) la suite des fonctions propres associées. Cependant, au lieu d'effectuer l'analyse spectrale des opérateurs Γ_X et $\Gamma_{X,n}$, on peut, comme dans [5] et [1], considérer la matrice $\mathbf{M}$ définie par

$$M_{i_1 i_2} = \frac{1}{n} \langle X_{i_1}, X_{i_2}\rangle, \tag{4}$$

pour $i_1, i_2 = 1, \dots, n$. Les valeurs propres non nulles de $\Gamma_{X,n}$ et les valeurs propres de $\mathbf{M}$ sont les mêmes. Soit $\hat{\boldsymbol{p}}_r = (\hat{p}_{1r}, \dots, \hat{p}_{nr})^\tau$ le vecteur propre de $\mathbf{M}$ associé à la valeur propre $\hat{\lambda}_r$ et $\hat{\eta}_{ir} = \langle X_i, \hat{g}_r\rangle$, on a

$$\hat{\eta}_{ir} = \sqrt{\hat{\lambda}_r}\, \hat{p}_{ir}, \tag{5}$$

pour tout $i = 1, \dots, n$ et pour tout $r \geqslant 1$ tel que $\hat{\lambda}_r > 0$. On obtient les fonctions propres $\hat{g}_r$ par

$$\hat{g}_r = \frac{1}{\sqrt{\hat{\lambda}_r}} \sum_{i=1}^n \hat{p}_{ir} X_i = \frac{\sum_{i=1}^n \hat{\eta}_{ir} X_i}{\sum_{i=1}^n \hat{\eta}_{ir}^2}, \tag{6}$$

pour tout $r \geqslant 1$. Ainsi, on construit une estimation de la matrice $\mathbf{M}$ définie par (4) en utilisant les estimations $\widetilde{W}_i$ de X_i, pour $i = 1, \dots, n$. L'estimateur le plus naturel $\hat{\mathbf{M}}$ de $\mathbf{M}$ semble être la matrice de taille $n \times n$ et de terme général $\hat{M}_{i_1 i_2} = \frac{1}{n} \langle \widetilde{W}_{i_1}, \widetilde{W}_{i_2}\rangle$, pour $i_1, i_2 = 1, \dots, n$. Cependant, l'estimateur $\langle \widetilde{W}_i, \widetilde{W}_i\rangle$ contient le terme

$$\sum_{j_1=1}^p W_i(t_{j_1})^2 \int_0^1 K\left(\frac{t - t_{j_1}}{h}\right)^2 \left[\sum_{j_3=1}^p K\left(\frac{t - t_{j_3}}{h}\right)\right]^{-2} \mathrm{d}t,$$

qui produit un biais dans l'estimation de M_{ii}, et qui sera donc supprimé (voir [5] et [1] pour une technique analogue). De plus, comme on estime des produits scalaires entre des courbes, on choisit de prendre une largeur de fenêtre $h_{i_1 i_2}$ pour l'estimation du produit scalaire entre la paire de courbes $\{X_{i_1}, X_{i_2}\}$, pour $i_1, i_2 = 1, \ldots, n$. Notons que d'autres choix pourraient être envisagés (comme par exemple $h_{i_1 i_2} = h_{i_1} h_{i_2}$). Ainsi, on considère l'estimateur de la matrice $\mathbf{M}$ dont le terme général sera défini par

$$\hat{M}_{i_1 i_2} = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{j_1=1}^{p} \sum_{j_2=1}^{p} W_{i_1}(t_{j_1}) W_{i_2}(t_{j_2}) \int_0^1 \frac{K(\frac{t-t_{j_1}}{h_{i_1 i_2}}) K(\frac{t-t_{j_2}}{h_{i_1 i_2}})}{[\sum_{j_3=1}^{p} K(\frac{t-t_{j_3}}{h_{i_1 i_2}})]^2} \,\mathrm{d}t, & \text{si } i_1 \neq i_2, \\ \dfrac{1}{n} \displaystyle\sum_{j_1=1}^{p} \sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{p} W_{i_1}(t_{j_1}) W_{i_2}(t_{j_2}) \int_0^1 \frac{K(\frac{t-t_{j_1}}{h_{i_1 i_2}}) K(\frac{t-t_{j_2}}{h_{i_1 i_2}})}{[\sum_{j_3=1}^{p} K(\frac{t-t_{j_3}}{h_{i_1 i_2}})]^2} \,\mathrm{d}t, & \text{si } i_1 = i_2. \end{cases} \tag{7}$$

Avec cette estimation $\hat{\mathbf{M}}$ de $\mathbf{M}$, on détermine les valeurs propres $\hat{\ell}_r$ de $\hat{\mathbf{M}}$ et les vecteurs propres $\hat{\boldsymbol{q}}_r$ correspondants, pour $r = 1, \ldots, n$. On en déduit des estimations $\hat{\beta}_{ir}$ de $\hat{\eta}_{ir}$ et $\hat{v}_r$ de $\hat{g}_r$ en utilisant les relations (5) et (6) dans lesquelles on remplace $\hat{\lambda}_r$ (resp. $\hat{p}_{ir}$, X_i et $\hat{g}_r$) par $\hat{\ell}_r$ (resp. $\hat{q}_{ir}$, $\widetilde{W}_i$ et $\hat{v}_r$). Pour la suite, on suppose que, p.s., $\hat{\lambda}_1 > \cdots > \hat{\lambda}_L > 0$ et $\hat{\ell}_1 > \cdots > \hat{\ell}_L > 0$. L'estimateur de α (qui suppose que les X_i sont connus) utilisant les L premières composantes principales est défini (voir [3]), pour tout $t \in [0, 1]$, par

$$\hat{\alpha}_1(t) = \frac{1}{n} \sum_{r=1}^{L} \sum_{i=1}^{n} \frac{Y_i}{\hat{\lambda}_r} \langle X_i, \hat{g}_r \rangle \hat{g}_r(t). \tag{8}$$

En se basant sur cette écriture, notre estimateur de α sera alors défini par

$$\hat{\alpha}_2(t) = \frac{1}{n} \sum_{r=1}^{L} \sum_{i=1}^{n} \frac{Y_i}{\hat{\ell}_r} \hat{\beta}_{ir} \hat{v}_r(t). \tag{9}$$

3. Résultat de convergence

On considère les hypothèses suivantes (voir [1] pour des hypothèses analogues). On notera aussi (comme dans [3]) $a_1 = 1/(\lambda_1 - \lambda_2)$ et $a_r = 1/\min(\lambda_{r-1} - \lambda_r, \lambda_r - \lambda_{r+1})$ pour $r \geqslant 2$, ainsi que $S_L = \sum_{r=1}^{L} a_r$.

(H.1) Il existe $0 < C_1, C_2, C_3, C_4 < +\infty$ telles que les fonctions g_r vérifient

$$\sup_{r \geqslant 1} \sup_{t \in [0,1]} \left| g_r(t) \right| \leqslant C_1,$$

$$\sum_{r=1}^{+\infty} \sum_{s=1}^{+\infty} \mathbb{E}\big(\langle X, g_r \rangle^2 \langle X, g_s \rangle^2\big) \leqslant C_2 \quad \text{et} \quad \sup_{r \geqslant 1} \sum_{s=1}^{+\infty} \sum_{q=1}^{+\infty} \mathbb{E}\big(\langle X, g_r \rangle^2 \langle X, g_s \rangle \langle X, g_q \rangle\big) \leqslant C_3.$$

On suppose de plus que $\sup_{r \geqslant 1} \sup_{t \in [0,1]} |\hat{g}_r(t)| = O_{\mathbb{P}}(1)$.

(H.2) Les valeurs propres de Γ_X vérifient $\lambda_1 > \lambda_2 > \cdots > 0$.

(H.3) X est p.s. deux fois dérivable et il existe $0 < C_4, C_5 < +\infty$ telles que $\sup_{t \in [0,1]} \mathbb{E}(X'(t)^4) \leqslant C_4$ et $\sup_{t \in [0,1]} \mathbb{E}(X''(t)^4) \leqslant C_5$.

(H.4) Les Y_i vérifient (en proba) $\sup_{i=1,\ldots,n} |Y_i| \leqslant C_6$ avec $0 < C_6 < +\infty$.

Théorème 3.1. *En prenant $h_{i_1 i_2} \sim p^{-\zeta}$ avec $\zeta \in [1/4, 1/2[$ et en supposant que L est négligeable devant S_L et que n est inférieur à p ou de l'ordre de p, sous les hypothèses* (H.1)–(H.4)*, on a*

$$\sup_{t \in [0,1]} \left| \hat{\alpha}_1(t) - \hat{\alpha}_2(t) \right| = O_{\mathbb{P}}\left[\frac{1}{\lambda_L} \left(\frac{L}{\lambda_L n^{1/2} p^{1/2}} + \frac{S_L}{\lambda_L^{1/2} n^{1/2} p} + L h^2 + \frac{L}{n^{1/2} p^{1/2} h^{1/2}} + \frac{S_L n^{1/2}}{\lambda_L^{3/2} p} \right) \right].$$

Éléments de preuve. Soit $\hat{\alpha}_1(t) - \hat{\alpha}_2(t) = A_1 + A_2 + A_3$ où

$$A_1 = \frac{1}{n}\sum_{r=1}^{L}\sum_{i=1}^{n}\left(\frac{Y_i}{\hat{\lambda}_r} - \frac{Y_i}{\hat{\ell}_r}\right)\langle X_i, \hat{g}_r\rangle \hat{g}_r(t), \qquad A_2 = \frac{1}{n}\sum_{r=1}^{L}\sum_{i=1}^{n}\frac{Y_i}{\hat{\ell}_r}\big(\langle X_i, \hat{g}_r\rangle - \hat{\beta}_{ir}\big)\hat{g}_r(t)$$

et

$$A_3 = \frac{1}{n}\sum_{r=1}^{L}\sum_{i=1}^{n}\frac{Y_i}{\hat{\ell}_r}\hat{\beta}_{ir}\big(\hat{g}_r(t) - \hat{v}_r(t)\big).$$

Les résultats de [1] et [5] permettent de contrôler les différences $|\hat{\lambda}_r - \ell_r|$, $|\langle X_i, \hat{g}_r\rangle - \hat{\beta}_{ir}|$ et $|\hat{g}_r(t) - \hat{v}_r(t)|$ et conduisent à

$$A_1 = O_{\mathbb{P}}\big[\lambda_L^{-2}(Ln^{-1/2}p^{-1/2} + S_L p^{-1})\big],$$
$$A_2 = O_{\mathbb{P}}\big[\lambda_L^{-1}(L\lambda_L^{-1/2}n^{-1}p^{-1/2} + \lambda_L^{-1/2}S_L n^{-1/2}p^{-1} + S_L n^{-1}p^{-1})\big]$$

et

$$A_3 = O_{\mathbb{P}}\big[\lambda_L^{-1}(Lh^2 + Ln^{-1/2}p^{-1/2}h^{-1/2} + \lambda_L^{-1/2}S_L p^{-1/2} + L\lambda_L^{-3/2}p^{-1/2} + \lambda_L^{-3/2}S_L n^{1/2}p^{-1})\big],$$

et ce indépendamment de r et t, d'où le résultat. Remarquons que le fait de prendre $h_{i_1i_2}$ de la forme $p^{-\zeta}$ avec $\zeta \in [1/4, 1/2[$ permet d'obtenir $n|\hat{M}_{i_1i_2} - M_{i_1i_2}| = O_{\mathbb{P}}(p^{-1/2})$ lors de l'estimation de la matrice $\mathbf{M}$. Notons enfin que la condition que n est inférieur à p ou de l'ordre de p a été posée de façon à simplifier les ordres de grandeurs des vitesses obtenues. □

Corollaire 3.2. *En plus des hypothèses du théorème précédent, si on prend h de la forme* $p^{-2/5}$, *en supposant que* $np^{-1/2}\lambda_L^{-1/2} \to +\infty$, *alors la vitesse devient*

$$\sup_{t\in[0,1]}\big|\hat{\alpha}_1(t) - \hat{\alpha}_2(t)\big| = O_{\mathbb{P}}\left(\frac{S_L n^{1/2}}{\lambda_L^{5/2}p}\right).$$

On suppose alors que $S_L n^{1/2}\lambda_L^{-5/2}p^{-1} \to 0$. *On note* $\|.\|_{\Gamma_X}$ *la semi-norme définie par* $\|u\|_{\Gamma_X}^2 = \langle \Gamma_X u, u\rangle$ *pour tout* $u \in L^2([0,1])$. *En supposant que* $\mathbb{E}(\|X\|^4) < +\infty$, $n\lambda_L^4 \to +\infty$ *et* $n\lambda_L^2(S_L)^{-2} \to +\infty$, *les résultats de* [3] *conduisent alors à*

$$\|\hat{\alpha}_2 - \alpha\|_{\Gamma_X} \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

On remarque ainsi que, sous les hypothèses faites, l'estimateur construit reste consistant (malgré le bruit sur la variable explicative) relativement à la semi-norme induite par l'opérateur de covariance. L'obtention d'une vitesse de convergence reste une question ouverte.

Remerciements

Je remercie Hervé Cardot et Pascal Sarda, les membres du groupe de travail STAPH à Toulouse, ainsi que le rapporteur de cette note dont les remarques ont contribué à améliorer ce travail.

Références

[1] M. Benko, W. Härdle, A. Kneip, Common functional principal components, SFB 649 Economic Risk Discussion Paper, 2006-010, 2005.
[2] H. Cardot, C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators in functional linear regression with errors-in-variables, Computational Statistics and Data Analysis 51 (2007) 4832–4848.
[3] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statistic and Probability Letters 45 (1999) 11–22.
[4] W. Härdle, Smoothing Techniques with Implementation in S, Springer, New York, 1991.
[5] A. Kneip, K.J. Utikal, Inference for density families using functional principal component analysis, Journal of the American Statistical Association 96 (2001) 519–542.
[6] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, second ed., Springer, New York, 2005.