Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 340 (2005) 851–854
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Convergence de l'estimateur spline cubique de lissage dans un modèle de régression longitudinale avec erreur de type processus

David Degras, Roxane Jallet

LSTA, boîte courrier 158, 8A, université Paris 6, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 16 septembre 2004 ; accepté après révision le 22 mars 2005

Disponible sur Internet le 17 mai 2005

Présenté par Paul Deheuvels

Résumé

Cette Note porte sur l'estimation d'une fonction de régression régulière par les splines cubiques de lissage. Un ordre de convergence est établi pour les erreurs quadratiques moyennes discrétisée et intégrée (MDSE et MISE) de l'estimateur, quand le bruit dans les données est un processus aléatoire. ***Pour citer cet article : D. Degras, R. Jallet, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

Abstract

Asymptotics for the smoothing cubic spline estimate in a longitudinal regression model with random process noise. This Note deals with the estimation of a smooth regression function by natural cubic splines. A convergence rate is obtained for the estimate's mean discretized and integrated squared error (MDSE and MISE) with random process noise in the data. ***To cite this article: D. Degras, R. Jallet, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

1. Introduction

En régression non paramétrique les méthodes de lissage permettent un compromis explicite entre ajustement aux données et stabilité de l'estimateur (ou encore entre biais et variance) via le choix d'un paramètre de lissage. Les splines cubiques de lissage sont largement utilisés en pratique à cause de leurs propriétés minimax d'une part (Speckman [10]) et de leur implémentation facile d'autre part (Reinsch [8]). Leur convergence, connue dans le cas de données comportant un bruit indépendant et identiquement distribué (iid) (Craven et Wahba [4], Ragozin [7]), est étudiée ici dans un modèle plus général de données longitudinales où le bruit est un processus aléatoire (situation

Adresses e-mail : degras@ccr.jussieu.fr (D. Degras), rjallet@yahoo.fr (R. Jallet).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2005.04.011

courante en pratique, par exemple en présence de corrélation et/ou d'hétéroscédasticité dans les données). Notre travail est à rapprocher de celui de Cardot et Diack [3] pour les splines hybrides, et plus généralement, de celui de Cardot [2] pour l'analyse en composantes principales lissée, de Boularan et al. [1] et de Nuñez-Anton et al. [6] pour les estimateurs à noyaux. Nous donnons ici les vitesses de convergence des estimateurs splines cubiques de lissage pour les critères d'erreur quadratique moyenne MDSE et MISE. Les résultats, présentés dans un cadre simple (répartition uniforme et équilibrée des points d'observation), peuvent être étendus à un modèle plus général de données déséquilibrées.

2. Modèle et estimateur de lissage

Nous considérons le modèle :

$$X_i(t_j) = f(t_j) + \varepsilon_i(t_j), \quad 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p, \tag{1}$$

où les trajectoires X_i $(1 \leqslant i \leqslant n)$ sont observées aux points $t_j = \frac{j}{p+1}$ $(1 \leqslant j \leqslant p)$, f étant la fonction de régression (partie déterministe) du modèle et les processus ε_i $(1 \leqslant i \leqslant n)$ étant des réalisations indépendantes du processus aléatoire de bruit $\varepsilon = \{\varepsilon(t),\ t \in [0,1]\}$. On suppose de plus que :

(H1) la fonction f est de classe C^2 sur $[0,1]$;
(H2) le processus ε est centré ($\mathbb{E}(\varepsilon(t)) = 0, t \in [0,1]$) et de variance finie ($\sup_{t\in[0,1]} \mathbb{E}(\varepsilon^2(t)) < \infty$).

Pour construire l'estimateur de lissage nous utilisons la pénalité classique sur la dérivée seconde qui permet de contrôler la courbure des solutions. L'espace des estimateurs est alors l'espace de Sobolev :

$$W^2\big([0,1]\big) = \big\{g : [0,1] \to \mathbb{R}\ ;\ g \text{ et } g' \text{ absolument continues}\,;\, g'' \in L^2\big([0,1]\big)\big\}.$$

Notons $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ la moyenne des trajectoires. Pour une valeur donnée du coefficient de lissage $\alpha > 0$, l'estimateur de lissage noté $\hat{f}_\alpha$ est défini de manière unique dès que $p \geqslant 2$ (Craven et Wahba [4], Schoenberg [9]) par :

$$\hat{f}_\alpha = \underset{g\in W^2([0,1])}{\operatorname{argmin}} \left[\frac{1}{p}\sum_{j=1}^{p}\big(\overline{X}(t_j) - g(t_j)\big)^2 + \alpha \int_0^1 g''(t)^2\,\mathrm{d}t\right]. \tag{2}$$

La fonction $\hat{f}_\alpha$ est une fonction spline cubique naturelle (cf. Green et Silverman [5]). Nous précisons à présent les notations utilisées dans cet article.

Soient $\mathbf{Q}$ et $\mathbf{R}$ les matrices de dimensions respectives $p\times(p-2)$ et $(p-2)\times(p-2)$, de coefficients $q_{kl} = -2/h$ si $k = l+1,\ 1/h$ si $k = l$ ou $k = l+2,\ 0$ sinon et $r_{kl} = h/6$ si $|k-l| = 1,\ 2h/3$ si $k = l,\ 0$ sinon. Définissons alors les matrices de dimension $p\times p$ $\mathbf{K} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}'$ et $\mathbf{S}_\alpha = (\mathbf{I} + \alpha p\mathbf{K})^{-1}$.

Soient aussi $\overline{\mathbf{X}} = [\overline{X}(t_1), \ldots, \overline{X}(t_p)]'$ le vecteur des moyennes des observations et $\mathbf{F} = [f(t_1), \ldots, f(t_p)]'$ le vecteur des valeurs de f aux points t_j $(1 \leqslant j \leqslant p)$.

Notons enfin $\mathbf{V} = [\mathrm{Cov}(\varepsilon(t_i), \varepsilon(t_j))]_{1\leqslant i,j\leqslant p}$ la matrice de variance–covariance associée au bruit ε et $N = np$ le nombre total d'observations.

L'estimateur $\hat{f}_\alpha$ est entièrement déterminé par ses valeurs aux points t_j contenues dans le vecteur $\widehat{\mathbf{F}}_\alpha = [\hat{f}_\alpha(t_1), \ldots, \hat{f}_\alpha(t_p)]'$ (cf. [5]). Celui-ci vérifie :

$$\widehat{\mathbf{F}}_\alpha = \mathbf{S}_\alpha \overline{\mathbf{X}}. \tag{3}$$

3. Convergence

Nous énonçons un résultat de convergence en moyenne quadratique pour chacun des deux critères

$$\mathrm{MDSE}(\widehat{\mathbf{F}}_\alpha, \mathbf{F}) = \mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p}\big(\widehat{f_\alpha}(t_j) - f(t_j)\big)^2\right] \quad \text{et} \quad \mathrm{MISE}(\hat{f}_\alpha, f) = \mathbb{E}\left[\int_0^1 \big(\hat{f}_\alpha(t) - f(t)\big)^2 \,\mathrm{d}t\right].$$

Théorème 3.1. *Sous les hypothèses* (H1) *et* (H2)*, il existe des constantes c_1 et c_2 indépendantes de n , p et α telles que* :

$$\mathrm{MDSE}(\widehat{\mathbf{F}}_\alpha, \mathbf{F}) \leqslant c_1\alpha + \frac{c_2}{n}. \tag{4}$$

Théorème 3.2. *Sous les hypothèses* (H1) *et* (H2)*, il existe des constantes c_3, c_4 et c_5 indépendantes de n, p et α telles que* :

$$\mathrm{MISE}(\hat{f}_\alpha, f) \leqslant c_3\alpha + \frac{c_4}{p^4} + \frac{c_5}{n}. \tag{5}$$

Remarques.

(i) Pour obtenir la convergence de la MDSE vers zéro, il suffit que $\alpha = \alpha(n, p) \to 0$ quand $n \to \infty$. Le paramètre α peut ainsi tendre vers zéro arbitrairement vite, contrairement au cas classique du bruit iid (ou bruit blanc) où s'ajoute la contrainte $\alpha p^{1/4} \to \infty$. Cela provient du fait que pour un bruit processus, le terme de variance est seulement contrôlé par le nombre n de trajectoires (en accord avec Cardot et Diack [3]). Pour faire converger la MISE vers zéro, il faut rajouter aux conditions précédentes la condition $p \to \infty$ afin de maîtriser le terme de biais sur tout l'intervalle $[0, 1]$.

(ii) Dans l'asymptotique de la MISE, la vitesse optimale en $\mathcal{O}(n^{-1}) + \mathcal{O}(p^{-4})$ coïncide avec celle obtenue par Cardot et Diack ([3], Théorème 2.2 et Remarque 1) pour les splines hybrides. Notons par contre que le champ de leurs résultats ne contient pas les splines de lissage.

(iii) Dans le cas d'un bruit blanc, une adaptation mineure des démonstrations permet de retrouver les vitesses optimales usuelles en $\mathcal{O}(N^{-4/5})$ pour la MDSE et en $\mathcal{O}(N^{-4/5}) + \mathcal{O}(p^{-4})$ pour la MISE.

(iv) Les résultats précédents s'étendent à un modèle de données longitudinales où on observe chaque courbe X_i en p_i points $t_{i,j}$:

$$X_i(t_{i,j}) = f(t_{i,j}) + \varepsilon_i(t_{i,j}), \quad 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p_i,\ 0 \leqslant t_{i,1} < \cdots < t_{i,p_i} \leqslant 1.$$

Réordonnons les $t_{i,j}$ en p points distincts t_j avec n_j répétitions et posons $h_{\min} = \min_j\{t_{j+1} - t_j\}$, $h_{\max} = \max\{t_1; \max_j\{t_{j+1} - t_j\}; 1 - t_p\}$ et ici $N = \sum_{i=1}^{n} p_i$. Alors sous les conditions $\frac{\max_j\{n_j\}}{\min_j\{n_j\}} = \mathcal{O}(1)$, $\min_j\{n_j\} \to \infty$, $\alpha \to 0$ pour la MDSE avec en plus $\frac{h_{\max}}{h_{\min}} = \mathcal{O}(1)$, $h_{\max} \to 0$ pour la MISE (ces conditions impliquant $n \to \infty$ pour la MDSE et $n, p \to \infty$ pour la MISE), on obtient alors des vitesses en $\mathcal{O}(\alpha) + \mathcal{O}(\frac{p}{N})$ pour la MDSE et en $\mathcal{O}(\alpha) + \mathcal{O}(h_{\max}^4) + \mathcal{O}(\frac{1}{\min_j\{n_j\}})$ pour la MISE.

4. Eléments de démonstration

Dans le Théorème 3.1, on décompose d'abord la MDSE en somme du biais au carré et de la variance :

$$\mathrm{MDSE}(\widehat{\mathbf{F}}_\alpha, \mathbf{F}) = \frac{1}{p}\,\mathbf{F}'(\mathbf{S}_\alpha - \mathbf{I})^2\mathbf{F} + \frac{1}{N}\,\mathrm{Trace}(\mathbf{S}_\alpha^2\mathbf{V}).$$

En majorant le biais au carré comme dans Craven et Wahba [4], on obtient :

$$\frac{1}{p}\mathbf{F}'(\mathbf{S}_\alpha - \mathbf{I})^2\mathbf{F} \leqslant \alpha \int_0^1 f''(t)^2\,\mathrm{d}t.$$

Pour contrôler le terme de variance, on écrit d'abord $\mathrm{Trace}(\mathbf{S}_\alpha^2\mathbf{V}) \leqslant \|\mathbf{S}_\alpha\|^2\,\mathrm{Trace}(\mathbf{V})$ (algèbre matricielle classique) puis on remarque que $\|\mathbf{S}_\alpha\| = 1$ (cf. Utreras [11]). Exploitant ensuite l'hypothèse (H2), on peut affirmer que $\mathrm{Trace}(\mathbf{V}) \leqslant p \times \sup_{t\in[0,1]} \mathbb{E}(\varepsilon^2(t))$, ce qui permet de conclure la démonstration.

Pour le Théorème 3.2, on part de la même décomposition de l'erreur quadratique (ici MISE) que précédemment. Le biais au carré se majore comme dans Ragozin [7]. Il existe ainsi des constantes c_3 et c_4 indépendantes de α, n et p telles que :

$$\int_0^1 \big(\mathbb{E}\hat{f}_\alpha(t) - f(t)\big)^2\,\mathrm{d}t \;\leqslant\; c_3\alpha + \frac{c_4}{p^4}.$$

Dans l'étude du terme de variance, on montre en premier lieu qu'il existe une matrice $\mathbf{M}$ de taille $p \times p$ et une constante c indépendante de α, n et p vérifiant :

$$\mathbb{E}\left[\int_0^1 \big(\hat{f}_\alpha(t) - \mathbb{E}\hat{f}_\alpha(t)\big)^2\,\mathrm{d}t\right] = \mathbb{E}\big[(\widehat{\mathbf{F}}_\alpha - \mathbb{E}\widehat{\mathbf{F}}_\alpha)'\mathbf{M}(\widehat{\mathbf{F}}_\alpha - \mathbb{E}\widehat{\mathbf{F}}_\alpha)\big] \quad \text{et} \quad \|\mathbf{M}\| \leqslant c \times h.$$

Il vient $\mathbb{E}[(\widehat{\mathbf{F}}_\alpha(t) - \mathbb{E}\widehat{\mathbf{F}}_\alpha(t))'\mathbf{M}(\widehat{\mathbf{F}}_\alpha(t) - \mathbb{E}\widehat{\mathbf{F}}_\alpha(t))] = \frac{1}{n}\,\mathrm{Trace}(\mathbf{S}_\alpha\mathbf{V}\mathbf{S}_\alpha\mathbf{M}) \leqslant \frac{1}{n}\|\mathbf{S}_\alpha\|^2\|\mathbf{M}\|\,\mathrm{Trace}(\mathbf{V})$, ce qui, joint aux relations précédentes, permet d'achever la démonstration.

Références

[1] J. Boularan, L. Ferré, P. Vieu, A nonparametric model for unbalanced longitudinal data, with application to geophysical data, Comput. Statist. 10 (1995) 285–298.

[2] H. Cardot, Nonparametric estimation of smoothed principal components analysis of sampled noisy functions, J. Nonparametr. Statist. 12 (4) (2000) 503–538.

[3] H. Cardot, C.A.T. Diack, Convergence en moyenne quadratique de l'estimateur de la régression par splines hybrides, C. R. Acad. Sci. Paris, Ser. I 326 (1998) 615–618.

[4] P. Craven, G. Wahba, Optimal smoothing of noisy data with spline functions, Numer. Math. 31 (1979) 316–356.

[5] P.J. Green, B.W. Silverman, Nonparametric Regression and Generalized Linear Models, Chapman and Hall, London, 1994.

[6] V. Nuñez-Anton, J. Rodriguez-Poo, P. Vieu, Longitudinal data with nonstationnary errors: a nonparametric three-stages approach, Test 8 (1999) 201–231.

[7] D.L. Ragozin, Error bounds for derivatives estimates based on spline smoothing of exact or noisy data, J. Approx. Theory 27 (1983) 335–355.

[8] C. Reinsch, Smoothing by spline functions, Numer. Math. 140 (1967) 177–183.

[9] I.J. Schoenberg, Spline functions and the problem of graduation, Proc. Nat. Acad. Sci. USA 52 (4) (1964) 947–950.

[10] P. Speckman, Spline smoothing and optimal rates of convergence in nonparametric regression models, Ann. Statist. 13 (3) (1985) 970–983.

[11] F. Utreras, Natural spline functions, their associated eigenvalue problem, Numer. Math. 42 (1983) 107–117.