Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 339 (2004) 717–720
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique/Probabilités

Loi du logarithme itéré pour les composantes du modèle additif de régression

Mohammed Debbarh

L.S.T.A., université de Paris 6, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 2 juillet 2003 ; accepté après révision le 27 septembre 2004

Présenté par Paul Deheuvels

Résumé

Dans le cadre des modèles additifs de régression, cette Note établit la loi du logarithme itéré pour les estimateurs des composantes additives obtenues par la méthode d'intégration marginale. Nos résultats sont établis dans le cas de vecteurs aléatoires indépendants et identiquement distribués. ***Pour citer cet article : M. Debbarh, C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

Abstract

Law of iterated logarithm for additive regression model components. In the setting of the additive model of the regression function, we study the iterated logarithm law for this model components pertaining with the marginal integration estimation method. Our results are stated in the i.i.d. random vectors framework. ***To cite this article: M. Debbarh, C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

1. Introduction

Soit $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ une suite de copies indépendantes du vecteur aléatoire $(\mathbf{X}, Y)$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$, $d \geqslant 2$. On considère la fonction de régression de Y sachant $\{\mathbf{X} = \mathbf{x}\}$ définie, pour tout $\mathbf{x} \in \mathbb{R}^d$, par $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. Il est bien connu depuis le travail de Stone [6] que, pour $0 < p < \infty$, la vitesse de convergence optimale au sens de la norme L_p, d'un estimateur non paramétrique de m est de l'ordre de $O(n^{-k/2k+d})$ quand la fonction est supposée k fois différentiable. Il ressort alors que cette vitesse de convergence est une fonction décroissante de la dimension d des covariables. Ce problème est connu sous le nom de « fléau de la dimension ». Dans le but de réduire l'effet de la dimension sur la vitesse de convergence, Stone [7] a proposé d'utiliser les

Adresse e-mail : debbarh@ccr.jussieu.fr (M. Debbarh).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2004.09.028

modèles additifs ; il a particulièrement étudié le modèle non paramétrique de régression dans lequel la fonction de régression multivariée est écrite comme la somme de fonctions univariées,

$$m(\mathbf{x}) = \mu + \sum_{l=1}^{d} m_l(x_l).$$

Depuis lors, plusieurs méthodes d'estimation dans les modèles additifs de régression ont été proposées, nous citons la méthode reposant sur l'algorithme du backfitting (Hastie et Tibshirani [3]), la méthode de spline (Stone [7]) et la méthode d'intégration marginale (Newey [5], Tjøstheim et Auestar [8] et Linton et Nielsen [4]). Des résultats de consistence des estimateurs ont été établis dans les situations analysées pour les diverses méthodes utilisées ; nous citons aussi le résultat de normalité asymptotique établi par Camlong, Sarda et Vieu [1]. Dans cette note, nous établissons la loi du logarithme itéré ponctuelle pour les estimateurs des composantes du modèle additif. Les composantes du modèle additif sont estimées en utilisant la méthode d'intégration marginale (voir, par exemple, Newey [5] pour le détail de cette méthode).

Pour construire la $l$ème composante du modèle additif de régression relative à la méthode d'intégration marginale, nous introduisons quelques notations. Posons $\mathbf{x}_{-l} = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_d)$, $q_{-l}(\mathbf{x}_{-l}) = \prod_{j\neq l} q_j(x_j)$ et $q(\mathbf{x}) = \prod_{i=1}^{d} q_l(x_l)$ où $q_1, \ldots, q_d$ sont des fonctions de densité réelles. La fonction m_l est alors définie, pour tout $x_l \in \mathbb{R}$, par

$$m_l(x_l) = \int_{\mathbb{R}^{d-1}} m(\mathbf{x}) q_{-l}(\mathbf{x}_{-l})\,\mathrm{d}\mathbf{x}_{-l} - \int_{\mathbb{R}^d} m(\mathbf{x}) q(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{1}$$

Pour estimer la fonction m_l, il suffit d'estimer la fonction de régression m et de l'injecter dans la formule (1). Quand la densité marginale f des covariables $\mathbf{X}$ est connue, l'estimateur de la fonction m est donné par

$$\hat{m}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{f(\mathbf{X}_i)} \left(\prod_{l=1}^{d} \frac{1}{h_{l,n}} K_l\left(\frac{x_l - X_{i,l}}{h_{l,n}} \right) \right), \tag{2}$$

où $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})$, $(h_{l,n})_n$, $1 \leqslant l \leqslant d$, sont des suites de réels positifs tendant vers zéro et $K_1, \ldots, K_d$ sont des noyaux de convolution réels. Lorsque la densité f est inconnue, on l'estime par

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n} \right),$$

où $(h_n)_n$ est une suite de réels positifs et K est un noyau de convolution défini sur $\mathbb{R}^d$ et on l'injecte ensuite dans la formule (2). Dans la suite, on suppose que la fonction f est de dérivées partielles d'ordre k lipschitziennes. $\hat{m}_1$ désigne l'estimateur de m_1 lorsque la densité f est connue.

Les résultats concernant chacune des composantes additives s'obtiennent de façon parfaitement similaire, ce qui permet d'exposer seulement des détails relatifs à la première composante. Pour cet exposé, quelques notations supplémentaires sont nécessaires. Pour tout $\mathbf{x} \in \mathbb{R}^d$, notons

$$H(\mathbf{x}) = E\left(Y^2 | \mathbf{X} = \mathbf{x}\right)$$

et pour tout $x_1 \in \mathbb{R}$, soit

$$\sigma_1^2(x_1) = \frac{\int_{\mathbb{R}} K_1^2(\mathbf{u})\,\mathrm{d}\mathbf{u}}{f_1(x_1)} \int_{\mathbb{R}^{d-1}} H(\mathbf{x}) \frac{q_{-1}^2(\mathbf{x}_{-1})}{f(\mathbf{x}_{-1}|x_1)}\,\mathrm{d}\mathbf{x}_{-1},$$

où $f(\mathbf{x}_{-1}|x_1)$ est la densité conditionnelle de $\mathbf{X}_{-l} = (X_1, \ldots, X_{l-1}, X_{l+1}, \ldots, X_d)$ sachant $X_1 = x_1$.

2. Résultats

Nous donnons d'abord les hypothèses nécessaires pour établir nos résultats.

H$_1$: $h_n \to 0$ et $nh_n^d \to \infty$ quand $n \to \infty$.
H$_2$: K est un noyau de convolution à support compact, d'intégrale 1 et de moments nuls jusqu'à l'ordre k.
H$_3$: $(h_{l,n})_{n\geqslant 1}$, $1 \leqslant l \leqslant d$, sont des suites de réels positifs telles que

$$h_{l,n} \to 0, \quad nh_{l,n} \to +\infty, \quad \frac{\log(n)}{nh_{l,n}} \to 0, \quad \lim_{n\to\infty} \frac{nh_{1,n}h_{l,n}^{2k}}{\log\log(n)} = 0 \quad \text{et} \quad \lim_{n\to\infty} \frac{h_{l,n}\log(n)}{h_n^d \log\log(n)} = 0.$$

H$_4$: Les noyaux K_l, $1 \leqslant l \leqslant d$, sont bornés, à supports compacts, d'ordre $k \geqslant 1$ et d'intégrale égale à 1.
H$_5$: Les fonctions $q_l(x_l)$, $1 \leqslant l \leqslant d$, sont des densités bornées, de carrés intégrables et de k premières dérivées bornées.
H$_6$: Les densités f et f_l de $\mathbf{X}$ et X_l respectivement, sont bornées, à supports compacts et continues sur leurs supports. En outre, il existe des nombres b, B, b_l et B_l tels que

$$0 < b_l \leqslant f_l(x_l) \leqslant B_l < \infty \quad \text{et} \quad 0 < b \leqslant f(\mathbf{x}) \leqslant B < \infty.$$

H$_7$: Il existe $M > 0$ tel que, pour tout $i \geqslant 1$, on a $|Y_i| \leqslant M < \infty$.
H$_8$: Il existe une suite de réels positifs (w_n) telle que les conditions de stabilité suivantes

$$\int_{|u_1|\leqslant w_n} \left|\mathrm{d}K_1\left(\frac{x_1-u_1}{h_{1,n}}\right)\right| + \left[K_1\left(\frac{x_1-w_n}{h_{1,n}}\right) + K_1\left(\frac{x_1+w_n}{h_{1,n}}\right)\right] = o\left(\sqrt{\frac{nh_{1,n}\log\log(n)}{(\log(n))^4}}\right),$$

et

$$\sum_{n\geqslant 1} \frac{1}{h_{1,n}\log\log(n)} E\left\{K_1^2\left(\frac{x_1-X_{i,1}}{h_{1,n}}\right) I_{\{|X_{i,1}|\geqslant w_n\}}\right\} < \infty,$$

où I_A est la fonction indicatrice de l'ensemble A, soient simultanément vérifiées.

Théorème 2.1. *Sous les conditions* **H**$_1$–**H**$_8$, *on a*

$$\limsup_{n\to+\infty} \sqrt{\frac{nh_{1,n}}{2\log\log(n)}}\big(\hat{m}_1(x_1) - E\big(\hat{m}_1(x_1)\big)\big) = \sqrt{\sigma_1^2(x_1)} + \int_{\mathbb{R}} \sqrt{\sigma_1^2(s)}q(s)\,\mathrm{d}s \quad p.s.$$

Corollaire 2.2. *On suppose les conditions* **H**$_1$–**H**$_8$ *satisfaites. Si en outre, pour un* $k \geqslant 1$, m *est de classe* C^k *et si pour tout* l, $1 \leqslant l \leqslant d$, $\|\frac{\partial^k m}{\partial v_l^k}\|_\infty < \infty$ *, alors*

$$\limsup_{n\to+\infty} \sqrt{\frac{nh_{1,n}}{2\log\log(n)}}\big(\hat{m}_1(x_1) - m_1(x_1)\big) = \sqrt{\sigma_1^2(x_1)} + \int_{\mathbb{R}} \sqrt{\sigma_1^2(s)}q(s)\,\mathrm{d}s \quad p.s.$$

Remarque 1. La normalisation obtenue dans nos résultats correspond à celle établie pour la régression unidimensionnelle (voir, par exemple, Hall [2]).

Éléments de preuve. Pour démonter le Théorème 2.1, on traite d'abord le cas où la densité f des covariables $\mathbf{X}$ est connue. L'utilisation de la décomposition

$$\frac{1}{\hat{f}_n} = \frac{1}{f} - \frac{\hat{f}_n - f}{\hat{f}_n f}$$

nous permet alors de traiter le cas où la densité f est inconnue.

L'estimation par la méthode d'intégration marginale nous permet d'écrire l'estimateur de la $l$ème composante du modèle additif sous la forme d'un processus de sommes partielles de variables aléatoires i.i.d. pour lequel on établit la loi du logarithme itéré en utilisant les arguments usuels pour les tableaux triangulaires (voir, Hall [2]).

Les conditions de stabilité nécessaires à ce résultat sont obtenues en utilisant l'approximation du processus empirique bidimensionnel par un pont brownien (voir, Tusnády [9]).

Références

[1] C. Camlong, P. Sarda, Ph. Vieu, Additive time series: the kernel integration method, Math. Methods Statist. 9 (1999) 358–375.
[2] P. Hall, Laws of iterated logarithm for nonparametric estimators, Z. Wahrsch. verw. Gebiete 59 (1981) 47–61.
[3] T.J. Hastie, R. Tibshirani, Generalized Additive Models, Chapman and Hall, London, 1990.
[4] O.B. Linton, J.B. Nielsen, A kernel method of estimating structured nonparametric regression based on marginal integration, Biometrika 82 (1995) 93–100.
[5] W.K. Newey, Kernel estimation of partial means, Econometric Theory 10 (1994) 233–253.
[6] C.J. Stone, Optimal global rates of convergences for non parametric regression, Ann. Statist. 10 (1982) 1040–1053.
[7] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 698–705.
[8] D. Tjøstheim, B. Auestar, Nonparametric identification of nonlinear time series: projections, J. Amer. Statist. Assoc. 89 (1994) 1398–1409.
[9] G. Tusnády, A remark on the approximation of the sample df in the multidimentional case, Period. Math. Hungar. 8 (1977) 53–55.