Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 345 (2007) 97–100
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Probabilités/Statistique

Convergence uniforme d'un estimateur de la fonction de régression additive en données censurées

Mohammed Debbarh [a], Vivian Viallon [a,b]

[a] *L.S.T.A., université de Paris 6, 175, rue du Chevaleret, 75013 Paris, France*
[b] *Unité de biostatistique, hôpital Cochin, faculté de médecine, Université Paris-Descartes, 75014 Paris, France*

Reçu le 9 octobre 2006 ; accepté après révision le 31 mai 2007

Disponible sur Internet le 13 juillet 2007

Présenté par Paul Deheuvels

Résumé

Dans cette Note, nous proposons d'établir la vitesse de convergence presque sûre optimale d'un estimateur de la fonction de régression additive en données censurées. Pour construire nos estimateurs, nous utilisons la méthode d'intégration marginale associée à un estimateur de type Inverse Probability of Censoring Weighted *[I.P.C.W.]*. ***Pour citer cet article : M. Debbarh, V. Viallon, C. R. Acad. Sci. Paris, Ser. I 345 (2007).***

Abstract

Uniform convergence of the estimator of an additive regression function under random censorship. In this Note, we establish the optimal almost sure rate of convergence for an estimator of the additive regression function under random censorship. To build our estimator, we used the method of marginal integration coupled with an Inverse Probability of Censoring Weighted *[I.P.C.W.]* estimate. ***To cite this article: M. Debbarh, V. Viallon, C. R. Acad. Sci. Paris, Ser. I 345 (2007).***

1. Introduction

Les données censurées interviennent dans de nombreux domaines d'application de la statistique, notamment en épidémiologie où les variables d'intérêt rencontrées (typiquement la survenue d'une maladie) sont liées à de multiples facteurs. Considérons par exemple la construction de score de risque pour le cancer du sein. Il s'agit, dans ce cas, d'évaluer la probabilité qu'une femme donnée a de développer cette pathologie dans un délai spécifié, conditionnellement à un certain nombre de facteurs de risque (entre cinq et quinze généralement). Dans ce cadre, l'utilisation d'estimateurs non paramétriques se heurte en pratique au problème bien connu du *fléau de la dimension*. En effet, les vitesses de convergence des estimateurs, habituellement utilisés pour évaluer la distribution de survie conditionnelle, dépendent fortement, en général, de la dimension des covariables (voir, par exemple, Dabrowska [5] ou Deheuvels et Derzko [7]).

Adresses e-mail : debbarh@ccr.jussieu.fr (M. Debbarh), viallon@ccr.jussieu.fr (V. Viallon).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.05.025

Pour répondre à ce problème, nous nous placerons, dans ce qui suit, sous les hypothèses du *modèle additif de régression*. Considérons le triplet aléatoire $(Y, C, \mathbf{X})$, à valeurs dans $\mathbb{R}\times\mathbb{R}\times\mathbb{R}^d$, $d \geqslant 2$. Ici, Y est la variable d'intérêt, C une variable de censure, et $\mathbf{X} = (X_1, \ldots, X_d)$ une variable concomitante, de conditionnement. Soit par ailleurs une fonction ψ mesurable. Nous nous intéressons à la fonction de régression de $\psi(Y)$ évaluée en $\mathbf{X} = \mathbf{x}$,

$$m_\psi(\mathbf{x}) = \mathbb{E}\big(\psi(Y) \mid \mathbf{X} = \mathbf{x}\big) =: \mu + \sum_{\ell=1}^{d} m_\ell(x_\ell), \quad \forall \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{1}$$

Pour assurer l'identifiabilité de (1), nous supposerons que $\mathbb{E}m_\ell(X_\ell) = 0$, $\ell = 1, \ldots, d$. Debbarh et Viallon [6] ont obtenu la vitesse de convegence en moyenne quadratique de l'estimateur de la régression additive *en données censurées* défini en (8) à partir de la méthode d'intégration marginale. Dans le même cadre, et sous des conditions plus générales sur la fonction ψ (voir **A**(ii) ci-après) nous permettant notamment de traiter le cas $\psi(y) = y$ et donc d'appliquer nos résultats à la fonction de régression *classique*, nous nous proposons d'établir la vitesse de convergence uniforme presque sûre. Ces vitesses de convergence sont identiques à celles obtenues par Camlong [1] et Camlong et al. [2] dans le cas non censuré.

2. Estimateur de la régression additive en données censurées

Nous travaillerons, par la suite, sur un échantillon aléatoire de taille $n \geqslant 1$, $(Y_i, C_i, \mathbf{X}_i)_{1\leqslant i\leqslant n}$, de triplets indépendants et de même loi que $(Y, C, \mathbf{X})$. Désignons par $\mathbb{1}_E$ la fonction indicatrice de E. Dans le cas de *censures à droite*, on observe les variables $Z_i = \min\{Y_i, C_i\}$, $\delta_i = \mathbb{1}_{\{Y_i\leqslant C_i\}}$ et $\mathbf{X}_i$. Introduisons les quantités $\overline{F}(t) = P(Y > t)$, $\overline{G}(t) = P(C > t)$ et $\overline{H}(t) = P(Z > t)$. Pour toute fonction de survie $\overline{L}$ définie sur $\mathbb{R}$, posons $T_L = \sup\{t\colon \overline{L}(t) > 0\}$. Soient $\{h_n\}_{n\geqslant 1}$ et $\{h_{j,n}\}_{n\geqslant 1}$, $j = 1, 2$ des suites de constantes réelles positives. Nous définissons l'estimateur $\hat{f}_n$ de la densité f du vecteur aléatoire $\mathbf{X}$, pour tout $\mathbf{x} \in \mathbb{R}^d$ et $n \geqslant 1$, par

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

où K est un *noyau* de convolution défini sur $\mathbb{R}^d$. Dans la suite de l'exposé, $\overline{G}_n^\star$ désigne l'estimateur de Kaplan–Meier [11] de $\overline{G}$. Pour tout $y \in \mathbb{R}$, on a, avec les conventions $\prod_\varnothing = 1$ et $0^0 = 1$,

$$\overline{G}_n^\star(y) = \prod_{1\leqslant i\leqslant n}\left(\frac{N_n(Z_i) - 1}{N_n(Z_i)}\right)^{\beta_i}, \quad \text{où } \beta_i = \mathbb{1}_{\{Z_i\leqslant y\}}(1-\delta_i), \text{ et } N_n(x) = \sum_{i=1}^{n}\mathbb{1}_{\{Z_i\geqslant x\}}. \tag{2}$$

Soient maintenant K_1, K_2 et K_3, des noyaux définis, respectivement, sur $\mathbb{R}$, $\mathbb{R}^{d-1}$ et $\mathbb{R}^d$. Posons, pour tout $\mathbf{x} = (x_1, \ldots, x_d)$, et pour tout $\ell = 1, \ldots, d$, $\mathbf{x}_{-\ell} = (x_1, \ldots, x_{\ell-1}, x_{\ell+1}, \ldots, x_d) \in \mathbb{R}^{d-1}$. Pour estimer la fonction de régression multivariée $m_\psi(\mathbf{x})$, définie en (1), nous ferons usage des estimateurs suivants (voir Carbonez et al. [3], Kohler et al. [12] et Jones et al. [10]). Adoptant la convention $0/0 = 0$, soit

$$\widetilde{m}^\star_{\psi,n,\ell}(\mathbf{x}) := \sum_{i=1}^{n} W^\ell_{n,i}(\mathbf{x})\frac{\delta_i\psi(Z_i)}{\overline{G}_n^\star(Z_i)}, \quad \text{où } W^\ell_{n,i}(\mathbf{x}) = \frac{K_1(\frac{x_\ell - X_{i,\ell}}{h_{1,n}})K_2(\frac{\mathbf{x}_{-\ell} - \mathbf{X}_{i,-\ell}}{h_{2,n}})}{nh_{1,n}h_{2,n}^{d-1}\hat{f}_n(\mathbf{X}_i)}, \tag{3}$$

pour tout $\ell = 1, \ldots, d$. On considérera également la quantité $\widetilde{m}^\star_{\psi,n}$, définie par

$$\widetilde{m}^\star_{\psi,n}(\mathbf{x}) := \sum_{i=1}^{n} W_{n,i}(\mathbf{x})\frac{\delta_i\psi(Z_i)}{\overline{G}_n^\star(Z_i)}, \quad \text{où, pour } i = 1, \ldots, n,\ W_{n,i}(\mathbf{x}) = \frac{K_3(\frac{\mathbf{x} - \mathbf{X}_i}{h_{1,n}})}{nh_{1,n}^d\hat{f}_n(\mathbf{X}_i)}. \tag{4}$$

Pour estimer les composantes additives de m_ψ, nous utilisons la méthode dite d'*intégration marginale*. Etant données $q_1, \ldots, q_d$, d densités réelles bornées, nous posons $q(\mathbf{x}) = \prod_{\ell=1}^d q_\ell(x_\ell)$ et $q_{-\ell}(\mathbf{x}_{-\ell}) = \prod_{j\neq\ell} q_j(x_j)$. Définissons alors, pour $\ell = 1, \ldots, d$,

$$\eta_\ell(x_\ell) = \int_{\mathbb{R}^{d-1}} m_\psi(\mathbf{x})q_{-\ell}(\mathbf{x}_{-\ell})\,\mathrm{d}\mathbf{x}_{-\ell} - \int_{\mathbb{R}^d} m_\psi(\mathbf{x})q(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{5}$$

Les fonctions η_ℓ, $\ell = 1, \ldots, d$, ainsi définies sont des composantes additives vérifiant les relations

$$\eta_\ell(x_\ell) = m_\ell(x_\ell) - \int_{\mathbb{R}} m_\ell(z_\ell) q_\ell(z_\ell)\,\mathrm{d}z_\ell \quad \text{et} \quad m_\psi(\mathbf{x}) = \sum_{\ell=1}^{d} \eta_\ell(x_\ell) + \int_{\mathbb{R}^d} m_\psi(\mathbf{z}) q(\mathbf{z})\,\mathrm{d}\mathbf{z}. \tag{6}$$

Considérant les expressions (4) et (5), un estimateur naturel de η_ℓ, pour $\ell = 1, \ldots, d$, est donné par

$$\hat{\eta}_\ell^\star(x_\ell) = \int_{\mathbb{R}^{d-1}} \widetilde{m}^\star_{\psi,n,\ell}(\mathbf{x}) q_{-\ell}(\mathbf{x}_{-\ell})\,\mathrm{d}\mathbf{x}_{-\ell} - \int_{\mathbb{R}^d} \widetilde{m}^\star_{\psi,n,\ell}(\mathbf{x}) q(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{7}$$

à partir duquel nous déduisons l'estimateur $\widehat{m}^\star_{\psi,add}(\mathbf{x})$ de la fonction de régression additive $m_\psi(\mathbf{x})$,

$$\widehat{m}^\star_{\psi,add}(\mathbf{x}) = \sum_{\ell=1}^{d} \hat{\eta}_\ell^\star(x_\ell) + \int_{\mathbb{R}^d} \widetilde{m}^\star_{\psi,n}(\mathbf{x}) q(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{8}$$

3. Présentation des hypothèses et résultats

Introduisons tout d'abord les hypothèses sur le triplet $(Y, C, \mathbf{X})$ qui seront utilisées pour l'obtention de nos résultats. Rappelons la définition (1) de m_ψ sous le modèle additif. En sus des hypothèses de base faites dans le §1, nous supposerons que, pour un entier $k \geqslant 1$ convenable,

(C.1) C et $(\mathbf{X}, Y)$ sont independants ;
(C.2) $\overline{G}$ est continue sur $\mathbb{R}$;
(C.3) il existe une constante M telle que, $\sup_t |\psi(t)| \leqslant M < \infty$;
(C.4) m_ψ est k-fois continument differentiable sur $\mathbb{R}^d$, et $\sup_{\mathbf{x}\in\mathbb{R}^d} |\frac{\partial^k}{\partial x_\ell^k} m_\psi(\mathbf{x})| < \infty$; $\ell = 1, \ldots, d$.

Soient $\mathcal{C}_1, \ldots, \mathcal{C}_d$, d parties compactes de $\mathbb{R}$, et soit $\mathcal{C} := \mathcal{C}_1 \times \cdots \times \mathcal{C}_d$ le compact produit de $\mathbb{R}^d$ correspondant. Pour toute partie $\mathcal{E}$ de $\mathbb{R}^q$, pour $q \geqslant 1$ quelconque, et pour tout $\alpha > 0$, définissons l'α-voisinage euclidien $\mathcal{E}^\alpha$ de $\mathcal{E}$, par $\mathcal{E}^\alpha = \{\mathbf{x}\colon \inf_{\mathbf{y}\in\mathcal{E}} \|\mathbf{x} - \mathbf{y}\|_{\mathbb{R}^q} \leqslant \alpha\}$, où $\|\cdot\|_{\mathbb{R}^q}$ désigne la norme euclidienne usuelle sur $\mathbb{R}^q$. Nous supposerons que les conditions suivantes sont satisfaites par les densités f et f_ℓ de $\mathbf{X}$ et X_ℓ, respectivement, pour $\ell = 1, \ldots, d$. On suppose que ces fonctions sont continues sur leurs domaines de définition, et qu'il existe une constante $\alpha > 0$ telle que les hypothèses (F.A-B) ci-dessous soient vérifiées.

(F.A) $\forall x_\ell \in \mathcal{C}_\ell^\alpha$, $f_\ell(x_\ell) > 0$, $\ell = 1, \ldots, d$, et $\forall \mathbf{x} \in \mathcal{C}^\alpha$, $f(\mathbf{x}) > 0$.
(F.B) f est k'-fois continument differentiable sur $\mathcal{C}^\alpha$ pour un entier $k' > dk$.

Les noyaux K, K_1, K_2 et K_3, utilisés dans le §2, et définis, respectivement, sur $\mathbb{R}^d$, $\mathbb{R}$, $\mathbb{R}^{d-1}$ et $\mathbb{R}^d$, seront supposés à supports compacts, continus et d'intégrales 1. Nous supposerons de plus que

(K.A) K_1 est lipschitzien ;
(K.B) les noyaux K_1 et K_3 sont d'ordre k, et le noyau K est d'ordre k'.

Nous travaillons avec des paramètres de lissage $h_n > 0$ et $h_{j,n} > 0$, $j = 1, 2$, pour $n = 1, 2, \ldots$, vérifiant les conditions (H.A-B) ci-dessous.

(H.A) $h_n = a_1 (\log n/n)^{1/(2k'+d)}$ pour un $0 < a_1 < \infty$.
(H.B) $h_{1,n} = a_2 (\log n/n)^{1/(2k+1)}$ pour un $0 < a_2 < \infty$, et $h_{2,n} = \mathrm{o}(1)$.

Enfin, nous nous placerons sous l'hypothèse (**A**), qui sera dite vérifié si l'une au moins des hypothèses (**A**)(i) ou (**A**)(ii) ci-dessous est vérifiée.

(**A**)(i) $\psi(y) = 0$ si $y \in (\tau, +\infty)$, avec $\tau < T_H$.
(**A**)(ii) $Y, C \in \mathbb{R}^+ \times \mathbb{R}^+$, $T_F < T_G$ et $\exists p \in (k/(2k+1), 1/2] : \int_0^{T_H} -\overline{F}^{-p/(1-p)}\, \mathrm{d}\overline{G} < \infty$.

Nous sommes, maintenant, en mesure d'énoncer notre résultat principal, dans le théorème ci-dessous :

Théorème 3.1. *On suppose que les conditions* (**A**)*,* (C.1-2-3-4), (F.A-B), (K.A-B) *et* (H.A-B) *sont satisfaites. Alors, on a, presque sûrement lorsque* $n \to \infty$*,*

$$\sup_{\mathbf{x}\in\mathcal{C}} \big|\widehat{m}^{\star}_{\psi,add}(\mathbf{x}) - m_{\psi}(\mathbf{x})\big| = \mathcal{O}\left(\left(\frac{\log n}{n}\right)^{\frac{k}{2k+1}}\right). \tag{9}$$

Éléments de preuve. Introduisons la fonction à valeurs réelles, $\Psi(y, c)$, définie pour $(y, c) \in \mathbb{R}^2$ par $\Psi(y, c) = [\mathbb{1}_{\{y\leqslant c\}}\psi(y \wedge c)]/\overline{G}(y \wedge c)$. L'hypothèse (C.1) nous permet d'adapter l'estimateur de la fonction de régression généralisée m_{Ψ} au cas des données censurées. En effet,

$$m_{\Psi}(\mathbf{X}) = \mathbb{E}\big(\Psi(Y, C)|\mathbf{X}\big) = \mathbb{E}\left\{\frac{\mathbb{1}_{\{Y\leqslant C\}}\psi(Z)}{\overline{G}(Z)}\Big|\mathbf{X}\right\} = \mathbb{E}\left\{\frac{\psi(Y)}{\overline{G}(Y)}\mathbb{E}[\mathbb{1}_{\{Y\leqslant C\}}|\mathbf{X}, Y]\Big|\mathbf{X}\right\} = m_{\psi}(\mathbf{X}).$$

Soit $\widehat{m}_{\psi,add}(\mathbf{x})$, l'estimateur de la fonction de régression additive m_{ψ} dans le cas où la fonction $\overline{G}$ est connue. En procédant comme dans [6], il est aisé de montrer que la version du Théorème 3.1 associée à $\widehat{m}_{\psi,add}(\mathbf{x})$ découle des mêmes arguments que ceux utilisés par Camlong [1] dans le cas non censuré. Pour traiter le cas où G est inconnue, il suffit d'évaluer l'écart entre $\widehat{m}^{\star}_{\psi,add}(\mathbf{x})$ et $\widehat{m}_{\psi,add}(\mathbf{x})$. En effet, sous la condition (**A**)(i), on obtient (voir [6] pour plus de détails)

$$\sup_{\mathbf{x}\in I} \big|\widehat{m}^{\star}_{\psi,add}(\mathbf{x}) - \widehat{m}_{\psi,add}(\mathbf{x})\big| = \mathcal{O}\Big(\sup_{t\leqslant\tau} \big|\overline{G}^{\star}_n(t) - \overline{G}(t)\big|\Big) \quad \text{p.s.} \tag{10}$$

La loi du logarithme itéré de Földes et Rejtő [8] nous permet alors de conclure.

Sous la condition (**A**)(ii), on obtient

$$\sup_{\mathbf{x}\in I} \big|\widehat{m}^{\star}_{\psi,add}(\mathbf{x}) - \widehat{m}_{\psi,add}(\mathbf{x})\big| = \mathcal{O}\Big(\sup_{-\infty<t\leqslant\mathcal{Z}_n} \big[\overline{G}^{\star}_n(t)\overline{G}(t)\big]^{-1} \sup_{t\leqslant T_F} \big|\overline{G}^{\star}_n(t) - \overline{G}(t)\big|\Big),$$

où $\mathcal{Z}_n = \max\{Z_i\colon\ \delta_i = 1, i = 1, \ldots, n\}$. Or, sous (**A**)(ii), il existe presque sûrement un n_0 tel que pour tout $n \geqslant n_0$, $\sup_{-\infty<t\leqslant\mathcal{Z}_n}[\overline{G}^{\star}_n(t)\overline{G}(t)]^{-1} < \infty$. Le résultat de Chen et Lo [4] (lorsque $p < 1/2$) et celui de Gu et Lai [9] (lorsque $p = 1/2$) permettent alors de conclure. □

Références

[1] C. Camlong, Convergence presque sûre de l'estimateur à noyau d'une fonction de régression additive sous une hypothèse de mélange, C. R. Acad. Sci. Paris, Sér. I 329 (1) (1999) 75–78.
[2] C. Camlong-Viot, P. Sarda, P. Vieu, Additive time series: the kernel integration method, Math. Methods Statist. 9 (4) (2000) 358–375.
[3] A. Carbonez, L. Györfi, E.C. van der Meulen, Partitioning-estimates of a regression function under random censoring, Statist. Decisions 13 (1) (1995) 21–37.
[4] K. Chen, S.H. Lo, On the rate of uniform convergence of the product-limit estimator: strong and weak laws, Ann. Statist. 25 (3) (1997) 1050–1087.
[5] D.M. Dabrowska, Uniform consistency of the kernel conditional Kaplan–Meier estimate, Ann. Statist. 17 (3) (1989) 1157–1167.
[6] M. Debbarh, V. Viallon, Mean square convergence for an estimator of the additive regression function under random censorship, C. R. Acad. Sci. Paris, Ser. I 344 (3) (2007) 1205–1210.
[7] P. Deheuvels and G. Derzko, Nonparametric estimation of conditional lifetime distributions under random censorship, in: Advances in Statistical Methods for the Health Sciences: Applications to Cancer and AIDS Studies, Genome Sequence Analysis and Survival Analysis, Springer, New York, 2006.
[8] A. Földes, L. Rejtő, A lil type result for the product-limit estimator, Z. Wahrsch. Verw. Gebiete 56 (1981) 75–86.
[9] M. Gu, T.L. Lai, Functional laws of the iterated logarithm for the product-limit estimator of a distribution function under random censorship or truncation, Ann. Probab. 18 (1990) 160–189.
[10] M.C. Jones, S.J. Davies, B.U. Park, Versions of kernel-type regression estimators, J. Am. Statist. Assoc. 89 (1994) 825–832.
[11] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Am. Statist. Assoc. 53 (1958) 457–481.
[12] M. Kohler, K. Máthé, M. Pintér, Prediction from randomly right censored data, J. Multivariate Anal. 80 (1) (2002) 73–100.