



Statistics

Uniform law of the logarithm for the local linear estimator of the conditional distribution function

Loi uniforme du logarithme pour l'estimateur linéaire local de la fonction de répartition conditionnelle

Sandie Ferrigno^a, Myriam Maumy-Bertrand^b, Aurélie Muller^a

^a *Équipe probabilités et statistiques, Institut Elie-Cartan Nancy, Nancy-Université, CNRS, INRIA, boulevard des Aiguillettes, B.P. 239, 54506 Vandoeuvre-lès-Nancy cedex, France*

^b *Équipe de statistique, Institut de recherche de mathématiques avancées, Université de Strasbourg, 7, rue René-Descartes, 67084 Strasbourg cedex, France*

ARTICLE INFO

Article history:

Received 30 July 2009

Accepted after revision 5 August 2010

Available online 15 September 2010

Presented by Paul Deheuvels

ABSTRACT

In this Note, we first recall the results of the behaviour of the nonparametric estimator of the conditional distribution function which we can find in the literature. We establish exact rate of strong uniform consistency for the local linear estimator of the conditional distribution function. Our methods of proofs are based upon modern empirical process theory in the spirit of the results of Einmahl and Mason (2000) [5] and Deheuvels and Mason (2004) [3].

© 2010 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

R É S U M É

Dans cette Note, nous rappelons d'abord les résultats sur le comportement de l'estimateur non paramétrique de la fonction de répartition conditionnelle, que nous pouvons trouver dans la littérature. Puis, nous établissons la vitesse exacte de la consistance forte uniforme pour l'estimateur linéaire local de la fonction de répartition conditionnelle. Pour démontrer nos résultats, nous utiliserons les théories récentes faisant intervenir des processus tous dérivant du processus empirique dans l'esprit des articles de Einmahl et Mason (2000) [5] et de Deheuvels et Mason (2004) [3].

© 2010 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

Version française abrégée

Soient $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$, des couples aléatoires indépendants et identiquement distribués à valeurs dans $\mathbb{R} \times \mathbb{R}$. (X, Y) est supposé admettre une densité jointe $f_{X,Y}$ sur $\mathbb{R} \times \mathbb{R}$.

Depuis ces trente dernières années, la littérature afflue sur l'étude des propriétés de la fonction de régression conditionnelle, de la densité conditionnelle et plus récemment de l'espérance conditionnelle $m_\psi(x)$, définie par l'équation (1). Citons par exemple Nadaraya [11], Watson [13], Devroye [4], Collomb [2], Härdle [9]. Einmahl et Mason [5] et Deheuvels et Mason [3] ont étudié de façon approfondie la fonction m_ψ et ses propriétés. Nous utilisons les outils développés dans ces articles pour établir nos démonstrations.

E-mail addresses: sandie.ferrigno@iecn.u-nancy.fr (S. Ferrigno), myriam.maumy@math.unistra.fr (M. Maumy-Bertrand), aurelie.muller@iecn.u-nancy.fr (A. Muller).

Dans cette Note, nous étudions l'estimation non paramétrique de la fonction de répartition conditionnelle, définie par l'équation (2). Cette fonction est souvent utilisée en Statistique, comme par exemple, dans le domaine médical pour l'estimation des courbes de référence (voir Gannoun et al. [8]). Nous rappelons les principaux résultats sur l'estimateur à noyau de type Nadaraya–Watson de la fonction de répartition conditionnelle, et nous apportons de nouveaux résultats sur l'estimateur linéaire local de cette fonction.

Stute [12] a démontré la convergence presque sûre de l'estimateur à noyau de type Nadaraya–Watson vers la fonction de répartition conditionnelle. Plus récemment, Einmahl et Mason [5] ont établi, sous certaines conditions de régularité des fonctions de densité, et sous certaines hypothèses sur la fenêtre h_n , une loi uniforme du logarithme pour l'estimateur à noyau de la fonction de répartition conditionnelle. Par la suite, Einmahl et Mason [6] ont donné une version plus générale, à savoir le couple aléatoire (X, Y) à valeurs dans $\mathbb{R}^r \times \mathbb{R}$, avec $r \in \mathbb{N}^*$. D'autre part, Blondin [1] a établi un résultat dans le cadre multivarié, à savoir (X, Y) à valeurs dans $\mathbb{R}^r \times \mathbb{R}^d$, avec $r, d \in \mathbb{N}^*$. La difficulté majeure de ce résultat est que l'approche développée par Einmahl et Mason [5] et Deheuvels et Mason [3] n'est plus applicable dans le cadre multivarié. En effet, dans ces deux articles, les processus empiriques sont indexés par des classes de fonctions à valeurs réelles et non pas vectorielles.

Le biais asymptotique de l'estimateur de la fonction de répartition conditionnelle a été étudié par Fan et Gijbels [7] via les polynômes locaux. Mint El Mouvid [10] a proposé une méthode fondée sur les U -statistiques qui induisent des calculs assez lourds. Dans la seconde section, nous utilisons la théorie des processus empiriques afin de proposer une loi uniforme du logarithme pour l'estimateur linéaire local de la fonction de répartition conditionnelle dans le cas où (X, Y) est à valeurs dans $\mathbb{R} \times \mathbb{R}$.

1. Introduction and earlier results

Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$, be independent and identically distributed, $\mathbb{R} \times \mathbb{R}$ valued random vectors with common joint Lebesgue density denoted by $f_{X,Y}$ and marginal densities denoted respectively by f_X and f_Y . We now introduce the conditional expectation of $\psi(Y)$ given $X = x$ defined by:

$$m_\psi(x) = \mathbb{E}(\psi(Y)|X = x) = \frac{1}{f_X(x)} \int_{\mathbb{R}} \psi(y) f_{X,Y}(x, y) dy, \tag{1}$$

whenever this regression function is meaningful. Here and elsewhere, ψ denotes a specified measurable function, which is assumed to be bounded on each compact subinterval of \mathbb{R} . Because of numerous applications, the problem of estimating the function m_ψ , the density function f_X , the regression function $m_{\psi=\text{id}}$ has been the subject of considerable interest during the last decades, for example Nadaraya [11], Watson [13], Devroye [4], Collomb [2], Härdle [9], and specially we mention two articles, Einmahl and Mason [5] and Deheuvels and Mason [3] for two reasons: the first is that these articles study the function m_ψ and its properties and the second is that we use the tools which are developed in these article in order to establish our proofs. We now choose $\psi(y) = \mathbb{1}_{\{y \leq t\}}$ with $t \in \mathbb{R}$ arbitrary but fixed, and $\mathbb{1}$ the indicator function, so we obtain the conditional distribution function, for all $t \in \mathbb{R}$, defined by:

$$F(t|x) = \mathbb{E}(\mathbb{1}_{\{Y \leq t\}}|X = x) = \mathbb{P}(Y \leq t|X = x), \quad \text{for all } x \in \mathbb{R}. \tag{2}$$

In this Note, we only study the conditional distribution function and the nonparametric estimator associated to this function. For example, this function is used, in medical domain, for the estimation of the references curves (see Gannoun et al. [8]).

Introduce the Nadaraya–Watson (see Nadaraya [11] and Watson [13]) estimator of the conditional distribution function $F(t|x)$, defined by

$$\widehat{F}_n(t|x) = \frac{\sum_{i=1}^n \mathbb{1}_{\{Y_i \leq t\}} K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}, \tag{3}$$

where $K(\cdot)$ is a real-valued kernel function on \mathbb{R} and $(h_n)_{n \geq 1}$ is a non-random (bandwidth) sequence of positive constants satisfying some assumptions. The Nadaraya–Watson estimator and its properties have been first studied by Stute [12].

Einmahl and Mason [5] have determined, under mild regularity conditions on the joint and marginal density functions and under hypotheses on the bandwidth h_n , exact rates of uniform strong consistency of kernel-type function estimators and specially for the conditional distribution function when the random vector (X, Y) is in $\mathbb{R} \times \mathbb{R}$. We recall here this result:

Corollary 1.1. (See Corollary 2 in Einmahl and Mason [5].) *Let $I = [a, b]$ be a compact interval. Assume that $f_{X,Y}$ and f_X satisfy some regularity conditions and moreover that h_n satisfies $h_n \searrow 0, nh_n \nearrow +\infty, |\log h_n|/\log \log n \rightarrow +\infty$ and $nh_n/\log n \rightarrow +\infty$ as $n \rightarrow +\infty$. Then we have for any kernel defined in [5] with probability one*

$$\lim_{n \rightarrow +\infty} \frac{\sup_{t \in \mathbb{R}} \sup_{x \in I} \sqrt{2nh_n} |\widehat{F}_n(t|x) - F_n(t|x)|}{\sqrt{|\log h_n|}} = \|K\|_2 / \inf_{x \in I} \sqrt{f_X(x)} \tag{4}$$

where $F_n(t|x) = \mathbb{E}(K((x - X)/h_n)\mathbb{1}_{\{Y \leq t\}})/h_n f_n(x)$, $f_n(x) = \mathbb{E}(\widehat{f}_n(x))$ where $\widehat{f}_n(x)$ is the kernel density estimator and $\|K\|_2^2 = \int_{\mathbb{R}} K^2(s) ds$.

In a more recent article, Einmahl and Mason [6] have given an uniform in bandwidth consistency of kernel-type function estimators, in the case where (X, Y) is in $\mathbb{R}^r \times \mathbb{R}$, $r \in \mathbb{N}^*$, and specially for the estimator $\widehat{F}_n(t|x)$. Moreover, Blondin establishes in [1] a similar result of the Corollary 1.1 in the multivariate case, i.e. (X, Y) is in $\mathbb{R}^r \times \mathbb{R}^d$, $r, d \in \mathbb{N}^*$. The major difficulty is that the approach developed by Einmahl and Mason [5] and Deheuvels and Mason [3] has not been used in the multivariate case. The results on the empirical processes indexed by classes of functions are established only for classes of real-valued functions.

It is a well-known fact the asymptotic bias of the Nadaraya–Watson estimator has a bad form. To overcome this problem, there exists an alternative: the local polynomial techniques described in Fan and Gijbels [7]. However, this method implies heavy calculations. Here, we do not use U -statistics in our proofs, but the theory of the empirical processes.

2. Main result

We are mainly focused on the stochastic part of the usual deviation. Our next task will be to extend the preceding results to the local polynomial least squares smoothers. For ease of presentation, we restrict ourselves to the local linear least squares estimator. The general results will found elsewhere.

Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$, be independent and identically distributed random couples in $\mathbb{R} \times \mathbb{R}$. Now K denotes a real-valued kernel function defined on \mathbb{R} and $I = [a, b], J = [a', b'] \supset I$ are two fixed compacts of \mathbb{R} .

We make the following assumptions:

- (F.1) $f_{X,Y}$ is continuous on $J \times \mathbb{R}$ and f_X is continuous and strictly positive on J ;
- (F.2) $Y \mathbb{1}_{\{X \in J\}}$ is bounded on \mathbb{R} ;
- (K.1) K is right-continuous function with bounded variation on \mathbb{R} ;
- (K.2) K is compactly supported and $\int_{\mathbb{R}} K(u) du = 1$;
- (K.3) $\int_{\mathbb{R}} uK(u) du = 0$ and $\int_{\mathbb{R}} u^2K(u) du \neq 0$;
- (H.1) $h_n \searrow 0$ and $nh_n \nearrow +\infty$, as $n \rightarrow +\infty$;
- (H.2) $nh_n/\log n \rightarrow +\infty$, as $n \rightarrow +\infty$;
- (H.3) $\log(h_n^{-1})/\log \log n \rightarrow +\infty$, as $n \rightarrow +\infty$.

Our aim will be to establish the strong uniform consistency of the local linear estimator of the conditional distribution function, defined by

$$\widehat{F}_n^{LL}(t|x) = \frac{\widehat{f}_{n,2}(x)\widehat{r}_{n,0}(x, t) - \widehat{f}_{n,1}(x)\widehat{r}_{n,1}(x, t)}{\widehat{f}_{n,0}(x)\widehat{f}_{n,2}(x) - \widehat{f}_{n,1}^2(x)} \tag{5}$$

where

$$\widehat{f}_{n,j}(x) = \frac{1}{nh_n} \sum_{i=1}^n \left(\frac{x - X_i}{h_n}\right)^j K\left(\frac{x - X_i}{h_n}\right), \quad \text{for } j = 0, 1, 2 \tag{6}$$

and

$$\widehat{r}_{n,j}(x, t) = \frac{1}{nh_n} \sum_{i=1}^n \mathbb{1}_{\{Y_i \leq t\}} \left(\frac{x - X_i}{h_n}\right)^j K\left(\frac{x - X_i}{h_n}\right), \quad \text{for } j = 0, 1. \tag{7}$$

This estimator is better than the Nadaraya–Watson estimator when the design is random and has the favorable property to reproduce polynomial of degree 1. Precisely, the local linear estimator has a high minimax efficiency among all possible estimators, including nonlinear smoothers. For the centering terms, we set

$$\widetilde{\mathbb{E}}(\widehat{F}_n^{LL}(t|x)) = \frac{f_{n,2}(x)r_{n,0}(x, t) - f_{n,1}(x)r_{n,1}(x, t)}{f_{n,0}(x)f_{n,2}(x) - f_{n,1}^2(x)},$$

where $f_{n,j}(x) = \mathbb{E}(\widehat{f}_{n,j}(x))$ for $j = 0, 1, 2$ and $r_{n,j}(x, t) = \mathbb{E}(\widehat{r}_{n,j}(x, t))$ for $j = 0, 1$. We obtain the following uniform law of the logarithm concerning the local linear estimator of the conditional distribution function:

Theorem 2.1. *Under (F.1)–(F.2), (H.1)–(H.3) and (K.1)–(K.3) we have,*

$$\lim_{n \rightarrow +\infty} \sup_{x \in I} \sqrt{\frac{nh_n}{2 \log(h_n^{-1})}} |\widehat{F}_n^{LL}(t|x) - \widetilde{\mathbb{E}}(\widehat{F}_n^{LL}(t|x))| = \sigma_{F,t}(I), \quad \text{almost surely} \tag{8}$$

where

$$\sigma_{F,t}^2(I) = \sup_{x \in I} \left(\frac{\text{Var}(\mathbb{1}_{\{Y \leq t\}} | X = x)}{f_X(x)} \right) \int_{\mathbb{R}} K^2(u) \, du = \sup_{x \in I} \left(\frac{F(t|x)(1 - F(t|x))}{f_X(x)} \right) \int_{\mathbb{R}} K^2(u) \, du.$$

Moreover, we have

$$\lim_{n \rightarrow +\infty} \sup_{t \in \mathbb{R}} \sup_{x \in I} \sqrt{\frac{nh_n}{2 \log(h_n^{-1})}} |\widehat{F}_n^{LL}(t|x) - \widetilde{\mathbb{E}}(\widehat{F}_n^{LL}(t|x))| = \sigma_F(I), \quad \text{almost surely} \tag{9}$$

where

$$\sigma_F^2(I) = \sup_{x \in I} \left(\frac{\sup_{t \in \mathbb{R}} F(t|x)(1 - F(t|x))}{f_X(x)} \right) \int_{\mathbb{R}} K^2(u) \, du = \frac{\|K\|_2^2}{4 \inf_{x \in I} f_X(x)}.$$

3. Elements of proofs

The proof of our main result follows the same line and can be inferred with little effort from the proof of Theorem 1 in Einmahl and Mason [5]. In a first step, we introduce a general local empirical process. For any $j = 0, 1, 2$ and continuous real valued functions $c(\cdot)$ and $d(\cdot)$ on J , set for $x \in J$,

$$W_{n,j}(x, t) = \sum_{i=1}^n (c(x)\mathbb{1}_{\{Y_i \leq t\}} + d(x)) K_j \left(\frac{x - X_i}{h_n} \right) - n \mathbb{E} \left((c(x)\mathbb{1}_{\{Y \leq t\}} + d(x)) K_j \left(\frac{x - X}{h_n} \right) \right) \tag{10}$$

where $K_j(u) = u^j K(u)$, $u \in \mathbb{R}$.

Theorem 3.1. Under (F.1)–(F.2), (H.1)–(H.3), (K.1)–(K.3) we have,

$$\lim_{n \rightarrow +\infty} \sup_{t \in \mathbb{R}} \sup_{x \in I} \sqrt{\frac{1}{2nh_n \log(1/h_n)}} |W_{n,j}(x, t)| = \sigma_W(I), \quad \text{almost surely} \tag{11}$$

where

$$\sigma_W^2(I) = \sup_{t \in \mathbb{R}} \sup_{x \in I} \mathbb{E} (c(x)\mathbb{1}_{\{Y \leq t\}} + d(x) | X = x) f_X(x) \int_{\mathbb{R}} K_j(u)^2 \, du.$$

In a second step, the deviation $\widehat{F}_n^{LL}(t|x) - \widetilde{\mathbb{E}}(\widehat{F}_n^{LL}(t|x))$ can be asymptotically expressed as a linear function of the bivariate empirical process. Indeed, applying Theorem 3.1, with $c(x) = 0$, $d(x) = 1$, $j = 0, 1, 2$, and then with $c(x) = 1$, $d(x) = 0$ and $j = 0, 1$, we get that:

$$\left(\frac{nh_n}{2 \log(1/h_n)} \right)^{1/2} \sup_{x \in I} |\widehat{f}_{n,j}(x) - f_{n,j}(x)| = \mathcal{O}(1) \quad \text{a.s., } j = 0, 1, 2, \tag{12}$$

$$\left(\frac{nh_n}{2 \log(1/h_n)} \right)^{1/2} \sup_{t \in \mathbb{R}} \sup_{x \in I} |\widehat{r}_{n,j}(x, t) - r_{n,j}(x, t)| = \mathcal{O}(1) \quad \text{a.s., } j = 0, 1. \tag{13}$$

Then, via Bochner’s Lemma, we have

$$f_{n,j}(x) = f_X(x)\mu_j(K) + o(1), \quad \text{uniformly in } x \in I, \quad j = 0, 1, 2, \tag{14}$$

$$r_{n,j}(x, t) = r(x, t)\mu_j(K) + o(1), \quad \text{uniformly in } (x, t) \in I \times \mathbb{R}, \quad j = 0, 1, \tag{15}$$

where $\mu_j(K) = \int_{\mathbb{R}} K_j(u) \, du$ and $r(x, t) = \int_{\mathbb{R}} \mathbb{1}_{\{Y \leq t\}} f_{X,Y}(x, y) \, dy$.

We deduce that, uniformly in $(x, t) \in I \times \mathbb{R}$,

$$\widehat{F}_n^{LL}(t|x) - \widetilde{\mathbb{E}}(\widehat{F}_n^{LL}(t|x)) = \frac{1}{f_X(x)} \{ \widehat{r}_{n,0}(x, t) - r_{n,0}(x, t) - F(t|x)[\widehat{f}_{n,0}(x) - f_{n,0}(x)] \} + o \left(\frac{nh_n}{2 \log(1/h_n)} \right)^{-1/2}.$$

The right-hand side above is exactly the process $W_{n,0}(x, t)/nh_n$ where we have chosen $c(x) = 1/f_X(x)$ and $d(x) = -F(t|x)/f_X(x)$. By combining this result with Theorem 3.1, we prove Theorem 2.1.

References

- [1] D. Blondin, Lois limites uniformes et estimation non paramétrique de la régression, Thèse soutenue en 2004, Université de Paris 6, France.
- [2] G. Collomb, Estimation non-paramétrique de la régression: Revue bibliographique, *Int. Statist. Rev.* 49 (1981) 75–93.
- [3] P. Deheuvels, D.M. Mason, General asymptotic confidence bands based on kernel-type function estimators, *Statist. Infer. Stochastic Process.* 7 (3) (2004) 225–277.
- [4] L. Devroye, The uniform convergence of the Nadaraya–Watson regression function estimate, *Can. J. Statist.* 6 (1978) 179–191.
- [5] U. Einmahl, D.M. Mason, An empirical process approach to the uniform consistency of kernel-type function estimators, *J. Theoret. Probab.* 13 (1) (2000) 1–37.
- [6] U. Einmahl, D.M. Mason, Uniform in bandwidth consistency of kernel-type function estimators, *Ann. Stat.* 33 (3) (2005) 1380–1403.
- [7] J. Fan, I. Gijbels, *Local Polynomial Modelling and its Applications*, Monographs on Statistics and Applied Probability, vol. 66, Chapman and Hall, 1996.
- [8] A. Gannoun, S. Girard, C. Guinot, J. Saracco, References curves based on nonparametric quantile regression, *Statistics in Medicine* 21 (2002) 3119–3155.
- [9] W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.
- [10] M. Mint El Mouvid, Sur l'estimateur linéaire local de la fonction de répartition conditionnelle, Thèse soutenue en 2000, Université de Montpellier 2, France.
- [11] E.A. Nadaraya, On estimating regression, *Theor. Probab. Appl.* 9 (1964) 141–142.
- [12] W. Stute, On almost sure convergence of conditional empirical distribution functions, *Ann. Probab.* 14 (3) (1986) 891–901.
- [13] G.S. Watson, Smooth regression analysis, *Sankhyā Ser. A* 26 (1964) 359–372.