Certainty bands for the conditional cumulative distribution function and applications

Aurélie Muller-Gueudin, Sandie Ferrigno, Myriam Maumy-Bertrand

Certainty bands for the conditional cumulative distribution function and applications

Aurélie Muller-Gueudin [1], Sandie Ferrigno [1], Myriam Maumy-Bertrand [2]

[1] *IECL, INRIA, BIGS, 54500 Vandoeuvre-les-Nancy,*
aurelie.gueudin@univ-lorraine.fr, [2] *IRMA, 67000 Strasbourg*

Résumé. Nous étudions l'estimateur polynomial local de la fonction de répartition conditionnelle. Nous donnons un résultat de consistence uniforme de cet estimateur, puis nous en déduisons des bandes de confiance asymptotiques de cette fonction. En corollaires, nous pouvons obtenir des estimateurs et des bandes de confiance asymptotiques pour les quantiles et la fonction de regression. Nous illustrons nos résultats par des simulations.

Mots-clés. Fonction de répartition conditionnelle, estimation polynomiale locale, bandes de confiance asymptotiques, fonction de régression, quantiles.

Abstract. In this paper, we establish uniform asymptotic certainty bands for the conditional cumulative distribution function. To this aim, we give exact rate of strong uniform consistency for the local linear estimator of this function. The corollaries of this result are the asymptotic certainty bands for the quantiles and the regression function. We illustrate our results with simulations.

Keywords. Conditional cumulative distribution function, local polynomial estimator, uniform asymptotic certainty bands, regression function, quantiles.

1 Introduction and hypotheses

Consider (X, Y), a random vector defined in $\mathbb{R}\times\mathbb{R}$. Throughout, we work with a sample $\{(X_i, Y_i)_{1\leqslant i\leqslant n}\}$ of independent and identically replica of (X, Y). We will assume that (X, Y) [resp $.X$] has density function $f_{X,Y}$ [resp. f_X] with respect to the Lebesgue measure. In this paper, we will mostly focus on a non parametric estimator of the conditional cumulative distribution function (*cond-cdf*) of Y given $X = x$, defined by :

$$\forall t \in \mathbb{R}, \quad F(t|x) = \mathbb{E}\left(\mathbf{1}_{\{Y\leqslant t\}}|X = x\right) = \mathbb{P}\left(Y \leqslant t|X = x\right). \tag{1}$$

Let $I = [a, b], J = [a', b'] \supsetneq I$, two fixed compacts of $\mathbb{R}$.

(F.1) $f_{X,Y}$ is continuous on $J \times \mathbb{R}$ and f_X is continuous and strictly positive on J ;

(F.2) $Y\mathbf{1}_{\{X\in J\}}$ is almost surely bounded on $\mathbb{R}$.

K denotes a positive-valued kernel function defined on $\mathbb{R}$, fulfilling the conditions :

(K.1) K is a right-continuous function with bounded variation on $\mathbb{R}$;

(K.2) K is compactly supported and $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$;

(K.3) $\int_{\mathbb{R}} uK(u)\mathrm{d}u = 0$ and $\int_{\mathbb{R}} u^2K(u)\mathrm{d}u \neq 0$. We note : $||K||_2^2 = \int_{\mathbb{R}} K^2(u)du$.

Further, introduce the following assumptions on the non-random sequence $(h_n)_{n\geqslant 1}$:

(H.0) for all n, $0 < h_n < 1$;

(H.1) $h_n \to 0$, as $n \to +\infty$;

(H.2) $nh_n/\log n \to +\infty$, as $n \to +\infty$;

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

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

where $^{(1)}$ denotes the order 1 of the local polynomial estimator, and

$$\widehat{f}_{n,j}(x, h_n) = \frac{1}{nh_n}\sum_{i=1}^{n}\left(\frac{x - X_i}{h_n}\right)^j K\left(\frac{x - X_i}{h_n}\right), \text{ for } j = 0, 1, 2, \tag{3}$$

$$\widehat{r}_{n,j}(x, t, h_n) = \frac{1}{nh_n}\sum_{i=1}^{n}\mathbf{1}_{\{Y_i\leqslant t\}}\left(\frac{x - X_i}{h_n}\right)^j K\left(\frac{x - X_i}{h_n}\right), \text{ for } j = 0, 1. \tag{4}$$

Remarks :

1. The Nadaraya-Watson estimator $\widehat{F}_n^{(0)}(t, h_n|x)$ can be also written with the functions $\widehat{f}_{n,j}$ and $\widehat{r}_{n,j}$ as
$$\widehat{F}_n^{(0)}(t, h_n|x) = \frac{\widehat{r}_{n,0}(x, t, h_n)}{\widehat{f}_{n,0}(x, h_n)}.$$
It is the local polynomial estimator of order 0 of the conditional cumulative distribution function.
2. The estimator $\widehat{F}_n^{(1)}(t, h_n|x)$ is better than the Nadaraya-Watson estimator when the design is random and has the favorable property to reproduce polynomial of order 1. Precisely, the local linear estimator has a high minimax efficiency among all possible estimators, including nonlinear smoothers (see Fan and Gijbels [1]).
3. The local polynomial estimator can be generalized to the orders $p \geqslant 2$, but it is not very interesting to study $p \geqslant 3$, see Fan and Gijbels [1], pp. 20-22 and 77-80. The argument is that the mean square error increases with p.

Now, we study the consistency of the estimator $\widehat{F}_n^{(1)}(t, h_n|x)$ via the decomposition :

$$\widehat{F}_n^{(1)}(t, h_n|x) - F(t|x) = \underbrace{\widehat{F}_n^{(1)}(t, h_n|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t, h_n|x)\right)}_{(1)} + \underbrace{\widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t, h_n|x)\right) - F(t|x)}_{(2)}$$

where, following the ideas of Deheuvels and Mason (see [2]), the centering term is :

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

where $f_{n,j}(x,h_n) = \mathbb{E}\left(\widehat{f}_{n,j}(x,h_n)\right)$ and $r_{n,j}(x,t,h_n) = \mathbb{E}\left(\widehat{r}_{n,j}(x,h_n)\right)$ for $j = 0,1,2$.

The *random part* (1) is the object of our theorem given in the following Section. Under (F.1-2), (H.1) and (K.1-3), the *deterministic term* (2), so-called bias, converges uniformly to 0 over $(x,t) \in I \times \mathbb{R}$.

2 Uniform consistency of the local linear estimator

The uniform law of the logarithm concerning the local linear estimator of the *cond-cdf*, is given in Theorem 2.1 below.

Theorem 2.1 *Under (F.1-2), (H.0-2) and (K.1-3), we have the convergence in probability, as $n \to \infty$:*

$$\sup_{x\in I}\sqrt{\frac{nh_n}{\log(h_n^{-1})}}\left|\widehat{F}_n^{(1)}(t,h_n|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t,h_n|x)\right)\right| \to \sigma_{F,t}(I) \tag{5}$$

where $\sigma_{F,t}^2(I) = 2||K||_2^2 \sup_{x\in I}\left(\frac{F(t|x)(1-F(t|x))}{f_X(x)}\right)$.

Moreover, we have, as $n \to \infty$:

$$\sup_{t\in\mathbb{R}}\sup_{x\in I}\sqrt{\frac{nh_n}{\log(h_n^{-1})}}\left|\widehat{F}_n^{(1)}(t,h_n|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t,h_n|x)\right)\right| \to \sigma_F(I) \tag{6}$$

where

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

We introduce the following quantity $L_n(x) := \sqrt{\frac{2nh_n}{\|K\|_2^2\log(h_n^{-1})}\widehat{f}_{n,0}(x,h_n)}^{-1}$. Note that L_n tends uniformly on I to 0 as $n \to \infty$. At the end of Section 1, assumptions are listed under which $F(t|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t,h_n|x)\right)$ converges to 0 uniformly. But to obtain the following result, this bias needs to be of order $o\left(L_n(x)\right)$.

Proposition 2.2 *Under (F.1-2), (H.0-2) and (K.1-3), and if h_n is such that the bias term $\sup_{t\in\mathbb{R}}\sup_{x\in I}\{L_n(x)\}^{-1}\left|F(t|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t,h_n|x)\right)\right| \to 0$ then we have, in probability, as $n \to \infty$:*

$$\sup_{t\in\mathbb{R}}\sup_{x\in I}\left\{L_n(x)\right\}^{-1}\left|\widehat{F}_n^{(1)}(t,h_n|x) - F(t|x)\right| \to 1. \tag{7}$$

3 Uniform asymptotic certainty bands

We show now how the Proposition 2.2 can be used to construct uniform asymptotic certainty bands for $F(t|x)$, in the following sense. Under the assumptions of the Proposition 2.2, we have, for each $0 < \varepsilon < 1$, and as $n \to +\infty$:

$$\mathbb{P}\left\{F(t|x) \in \left[\widehat{F}_n^{(1)}(t, h_n|x) \pm (1+\varepsilon)L_n(x)\right], \text{ for all } (x,t) \in I \times \mathbb{R}\right\} \to 1 \tag{8}$$

and

$$\mathbb{P}\left\{F(t|x) \in \left[\widehat{F}_n^{(1)}(t, h_n|x) \pm (1-\varepsilon)L_n(x)\right], \text{ for all } (x,t) \in I \times \mathbb{R}\right\} \to 0. \tag{9}$$

Whenever (8) and (9) hold jointly for each $0 < \varepsilon < 1$, we have the following corollary :

Corollary 3.1 *Under (F.1-2), (H.0-2) and (K.1-3), and if h_n is such that the bias term $\sup_{t\in\mathbb{R}} \sup_{x\in I}\{L_n(x)\}^{-1}|F(t, h_n|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t, h_n|x)\right)| \to 0$ then the interval*

$$\left[\widehat{F}_n^{(1)}(t, h_n|x) \pm L_n(x)\right] \tag{10}$$

provides uniform asymptotic certainty bands (at an asymptotic confidence level of 100%) for the cond-cdf $F(t|x)$, uniformly in $(x,t) \in I \times \mathbb{R}$.

Let $m(x) = \mathbb{E}(Y|X = x)$ the regression function and $\widehat{m}_n^{(1)}(x) = \int y\widehat{F}_n^{(1)}(dy, h_n|x)$ its local linear estimator. The Proposition 2.2 has the following corollary.

Corollary 3.2 *Under (F.1-2), (H.0-2) and (K.1-3), and if h_n is such that the bias term $\sup_{t\in\mathbb{R}} \sup_{x\in I}\{L_n(x)\}^{-1}\left|F(t|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t, h_n|x)\right)\right| \to 0$ and the variable Y lives in the real interval $[\alpha, \beta]$, then the interval*

$$\left[\widehat{m}_n^{(1)}(x) \pm (\beta - \alpha)L_n(x)\right] \tag{11}$$

provides uniform asymptotic certainty bands (at an asymptotic confidence level of 100%) for the conditional regression function $m(x)$, uniformly in $x \in I$.

For $0 < \alpha < 1$, let the conditional α-quantile $q_\alpha(x) = \inf\{t \in \mathbb{R} : F(t|x) \geqslant \alpha\}$ and its local linear estimator $\widehat{q}_{\alpha,n}^{(1)}(x) = \inf\{t \in \mathbb{R} : \widehat{F}_n^{(1)}(t, h_n|x) \geqslant \alpha\}$. The Proposition 2.2 has the following corollary for the conditional quantiles.

Corollary 3.3 *Under (F.1-2), (H.0-2) and (K.1-3), if h_n is such that the bias term $\sup_{t\in\mathbb{R}} \sup_{x\in I}\{L_n(x)\}^{-1}\left|F(t|x) - \widehat{\mathbb{E}}\left(\widehat{F}_n^{(1)}(t, h_n|x)\right)\right| \to 0$ and if the function $x \mapsto f_{X,Y}\left(x, q_\alpha(x)\right) \neq 0$ for all $x \in I$, then the interval*

$$\left[\widehat{q}_{\alpha,n}^{(1)}(x) \pm \frac{2L_n(x)f_X(x)}{f_{X,Y}\left(x, q_\alpha(x)\right)}\right] \tag{12}$$

provides uniform asymptotic certainty bands (at an asymptotic confidence level of 100%) for the conditional α-quantile $q_\alpha(x)$, uniformly in $x \in I$.

The proofs of these results are available in our article [5].

4 A simulation study

We consider the model $Y = 2\sin(\pi X) + \epsilon$ where X, ϵ are independent random variables having a common distribution with density $1 - |x|$ on $[-1, 1]$. This is a model already studied by [3] and [4].

We worked with the sample size $n = 100$ and the Epanechnikov kernel. The bandwidth is selected by minimization of a cross-validation criteria (see [6]) :

$$CV(h, (X_i, Y_i)_{1 \leqslant i \leqslant n}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left(\mathbf{1}_{Y_i \leqslant Y_j} - \widehat{F}_{-i}^{(1)}(Y_j, h | X_i) \right)^2$$

where $\widehat{F}_{-i}^{(1)}(y, h|X_i)$ is defined by :

$$\frac{\sum_{j \neq i} \left(\frac{X_j - X_i}{h} \right)^2 K\left(\frac{X_j - X_i}{h} \right) \sum_{j \neq i} \mathbf{1}_{Y_j \leqslant y} K\left(\frac{X_j - X_i}{h} \right) - \sum_{j \neq i} \left(\frac{X_j - X_i}{h} \right) K\left(\frac{X_j - X_i}{h} \right) \sum_{j \neq i} \mathbf{1}_{Y_j \leqslant y} \frac{X_j - X_i}{h} K\left(\frac{X_j - X_i}{h} \right)}{\sum_{j \neq i} \left(\frac{X_j - X_i}{h} \right)^2 K\left(\frac{X_j - X_i}{h} \right) \sum_{j \neq i} K\left(\frac{X_j - X_i}{h} \right) - \left[\sum_{j \neq i} \left(\frac{X_j - X_i}{h} \right) K\left(\frac{X_j - X_i}{h} \right) \right]^2}$$

The Figure 1 gives the graph of a sample $(X_i, Y_i)_{i=1,\ldots,n}$, and its associated *cond-cdf*. The certainty bands are too narrow since $L_n(x)$ is very small. It is not satisfactory since the coverage probabilities is far from 1.

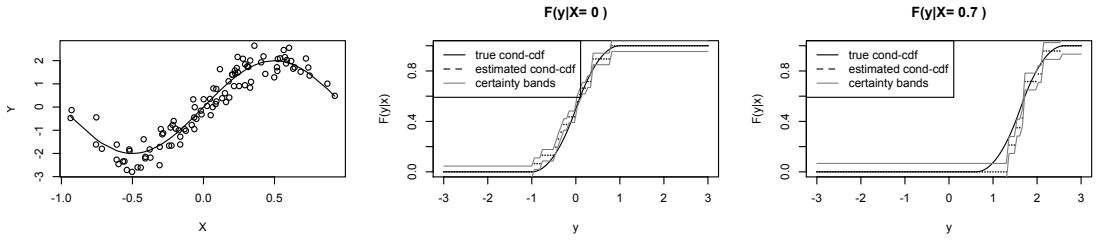

FIGURE 1 – Graph of a sample $(X_i, Y_i)_{i=1,\ldots,n}$ and corresponding estimations.

Another solution is to compute B bootstrapped samples from $(X_i, Y_i)_{i=1,\ldots,n}$ and to take the envelope of their B associated certainty bands (this envelope is named in the sequel bootstrapped certainty band). The figure 2 shows the bootstrapped certainty bands for $x = 0, 0.7$ and $B = 100$. The coverage probabilities of the bootstrapped certainty bands is estimated by $N_{rep} = 100$ repetitions of the data generation, and the results are illustrated in the figure 3. We used $B = 30$ and $B = 100$ (the optimal h is computed at each new data generation, i.e. N_{rep} times). When $B = 100$, the coverage probabilities are higher that 0.85.

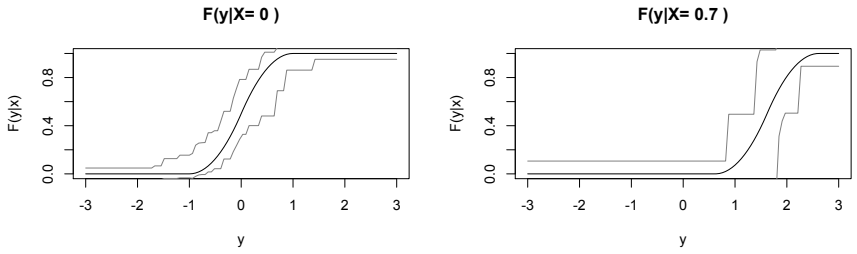

FIGURE 2 – Bootstrapped certainty bands of the *cond-cdf* for $x = 0$ (left) and $x = 0.7$ (right).

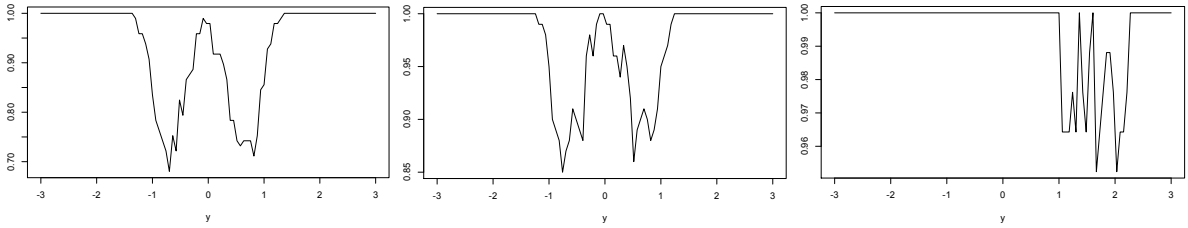

FIGURE 3 – Coverage probabilities of the true *cond-cdf* after $N_{rep} = 100$ repetitions, $x = 0, B = 30$ (left), $x = 0, B = 100$ (middle), and $x = 0.7, B = 100$ (right).

Bibliographie

[1] J. Fan, I. Gijbels, 1996, *Local polynomial modeling and its applications*. Monographs on Statistics and Applied Probability, Chapman and Hall, Vol. 66.
[2] P. Deheuvels, D.M. Mason, 2004, *General asymptotic confidence bands based on kernel-type function estimators*, Statist. Infer. Stochastic Process, 7(3), pp. 225-277.
[3] P. Hall, R.CL. Wolff, Q. Yao, 1999, *Methods for estimating a conditional distribution function*, JASA, 94 (445). pp. 154-163.
[4] N. Veraverbeke, I. Gijbels, M. Omelka, 2014, *Preadjusted non-parametric estimation of a conditional distribution function*, J.R. Statist. Soc. B, 76(2), pp. 399-438.
[5] S. Ferrigno, B. Foliguet, M. Maumy-Bertrand, A. Muller-Gueudin, 2014. *Certainty bands for the conditional cumulative distribution function and applications.* hal-01025443
[6] Q. Li, J. Lin, J.S. Racine, 2013, *Optimal bandwidth selection for nonparametric conditional distribution and quantile functions*, J. of business and eco. statist., 31(1), pp. 57-65.