STRONG CONSISTENCY OF A CONDITIONAL MODE ESTIMATOR IN THE PRESENCE OF DOUBLY CENSORED DATA

Hadjer Benchoulak[a], *Mohamed Boukeloua*[a,b] *and Fatiha Messaci*[a]

[a] Laboratoire LAMASD, Département de Mathématiques,
Université Frères Mentouri Constantine 1,
route d'Ain El Bey, 25017 Constantine, Algeria
Tel: 00 213 31 70 14 08

[b] Laboratoire LGPDDPS, Ecole Nationale Polytechnique de Constantine,
Constantine, Algeria
Tel: 00 213 31 78 51 68, Fax: 00 213 31 78 51 74
E-mail: boukeloua.mohamed@gmail.com

ABSTRACT

In this work, we study a double kernel conditional density and a conditional mode estimators when the response variable is doubly censored. We establish the uniform strong consistency of the conditional density estimator and its derivatives. Then, we deduce the uniform strong consistency of the conditional mode estimator that results. Finally, we illustrate the quality of the estimation through a simulation study.

1. INTRODUCTION

Let X be a real random variable (r.r.v.) and let Y be a nonnegative r.r.v. such that the density $f(x,y)$ of (X,Y) exists. Throughout this work f_X (resp. $f^{(0)}(.|x)$) stands for the marginal density of X (resp. the conditional density of Y given $X=x$). So, for any x such that $f_X(x) \neq 0$, we have $f^{(0)}(y|x) = \frac{f(x,y)}{f_X(x)}$. Moreover, for any r.r.v. U, $S_U(t) = P(U > t)$ denotes the survival function of U. We assume that Y is doubly censored, so we can only observe a sample of i.i.d. observations $(X_i, Z_i, \delta_i)_{1\leq i\leq n}$ of $(X, Z := \max(\min(Y,R),L), \delta)$, where R and L are nonnegative censoring r.r.v. such that $L \leq R$ a.s. and $\delta = \begin{cases} 1 & \text{if } L \leq Y \leq R, \\ 2 & \text{if } Y > R, \\ 3 & \text{if } Y < L, \end{cases}$
is the indicator of censorship. Our aim is to estimate $f^{(0)}(y|x)$ from the sample $(X_i, Z_i, \delta_i)_{1\leq i\leq n}$ and to deduce an estimator for the conditional mode on a compact set $S' \subset]0,\infty[$. We assume that the conditional mode $\theta(x)$ is unique on the compact S', it is defined by

$$\theta(x) = arg\max_{y\in S'} f^{(0)}(y|x).$$

We propose to estimate $f^{(0)}(y|x)$ by the following double kernel estimatoer.

$$f_n^{(0)}(y|x) := \frac{\sum_{i=1}^n \frac{1}{h_n g_n} K\left(\frac{x-X_i}{h_n}\right) H\left(\frac{y-Y_i}{g_n}\right) \frac{1_{\{\delta_i=1\}}}{|S_R^{(n)}(Z_i) - S_L^{(n)}(Z_i)| + V_n}}{\sum_{i=1}^n \frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)},$$

where K and H are appropriate kernel functions, h_n and g_n are two sequences of strictly positive real numbers tending to zero, the sequence $V_n := 1/n$ serves to avoid division by zero and $S_R^{(n)}$ (resp. $S_L^{(n)}$) is the self consistent estimator of S_R (resp. S_L) introduced in [(Turnbull, 1974)].
We also need to introduce for $p = 1,2$ the pth partial derivative with respect to y of $f_n^{(0)}(y|x)$, given by

$$f_n^{(p)}(y|x) := \frac{\sum_{i=1}^{n} \frac{1}{h_n g_n} K\left(\frac{x-X_i}{h_n}\right) H^{(p)}\left(\frac{y-Y_i}{g_n}\right) \frac{1_{\{\delta_i=1\}}}{|S_R^{(n)}(Z_i)-S_L^{(n)}(Z_i)|+V_n}}{\sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)},$$

where $H^{(p)}$ is the pth derivative of H.
On the basis of $f_n^{(0)}(y|x)$, a natural conditional mode estimator $\theta_n(x)$ of $\theta(x)$ is defined as follows.

$$\theta_n(x) = arg \max_{y \in S'} f_n^{(0)}(y|x).$$

2. HYPOTHESES AND RESULTS

To establish the uniform strong consistency of $f_n^{(p)}(y|x)$ ($p = 0,2$) and $\theta_n(x)$, we need the following hypotheses used in [(Chang and Yang, 1987)] and[(Chang, 1990)].

H_1 : S_Y, S_R and S_L are continuous functions of t for $t \geq 0$ and $0 < S_Y(t) < 1$ for $t > 0$.

H_2 : $0 < P(L < t < R) = S_R(t) - S_L(t)$ for every $t > 0$.

H_3 : (X,Y) and (L,R) are independent.

We also need the following assumptions which are standard in the nonparametric setting.

N_0 : The kernel K is a lipschitzian density with compact support.

N_1 : The kernel H is a twice continuously differentiable density with compact support and such that $H^{(2)}$ is lipschitzian.

N_2 : $\lim_{n\to\infty} h_n = 0, \lim_{n\to\infty} g_n = 0$ and $\lim_{n\to\infty} \frac{\ln n}{n h_n g_n^5} = 0$.

N_3 : The marginal density $f_X(x)$ of the variable X is continuously differentiable and verify $\inf_{x\in S} f_X(x) > 0$, where S is a compact set of $\mathbb{R}$.

N_4 : The joint density $f(.,.)$ is three times continuously differentiable.

Now we can state our claimed results.

Theorem 1 *Under assumptions $H_1 - H_3$ and $N_0 - N_4$, for $p = 0 - 2$, we have*

$$\lim_{n\to\infty} \sup_{x\in S} \sup_{y\in S'} |f_n^{(p)}(y|x) - f^{(p)}(y|x)| = 0 \quad a.s.$$

To deal with the conditional mode estimator, we need to add a standard hypothesis which stipulates the uniform uniqueness property of the conditional mode.

$$\begin{aligned} &\forall \alpha > 0,\ \exists b > 0 \text{ such that for any function } \eta : S \to \mathbb{R}, \text{we have} \\ &\sup_{x\in S} |\theta(x) - \eta(x)| \geq \alpha \Rightarrow \sup_{x\in S} |f^{(0)}(\theta(x)|x) - f^{(0)}(\eta(x)|x)| \geq b. \end{aligned} \tag{1}$$

Theorem 2 *Assume that* (1) *holds. Then, under assumptions $H_1 - H_3$ and $N_0 - N_4$, we have*

$$\lim_{n\to\infty} \sup_{x\in S} |\theta_n(x) - \theta(x)| = 0 \quad a.s.$$

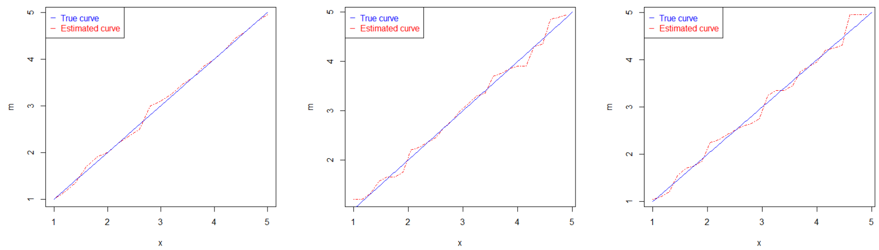

FIGURE 1 – **M1**, $n = 100$ and from left to right CR $= 0\%$, CR$\simeq 30\%$ and CR$\simeq 60\%$

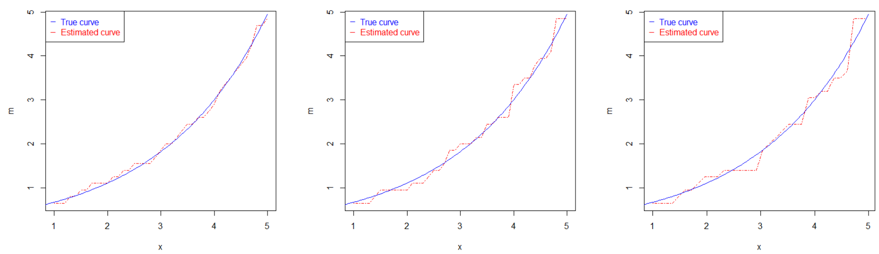

FIGURE 2 – **M2**, $n = 300$ and from left to right CR $= 0\%$, CR $\simeq 30\%$ and CR$\simeq 60\%$

3. SIMULATION STUDY

In this section, we examine the finite sample performance of the conditional mode estimator $\theta_n(x)$, for the usual following models.
M1 : $Y = X + 0.2\varepsilon$, **M2** : $Y = exp(0.5X - 0.9) + 0.2\varepsilon$ and **M3** : $Y = cos(1.5X) + 0.2\varepsilon$, where X and ε are independent and follow respectively, the exponential distribution with parameter $1/3$ and the standard normal distribution .
The right censoring variable is chosen as $R = L + V$, where the left censoring variable L and the variable V follow the exponential distribution with parameters λ_1 and λ_2 respectively. We take different values of λ_1 and λ_2 to obtain different censoring rates (CR). We choose the same Epanechnikov kernel for K and H and we take $h_n = g_n$. Finally, we use an optimal bandwidth obtained by minimizing the supremum norm between the estimated and the true values over a set of the bandwidths values on $[0, 0.3]$.
We first assess the effect of the censoring rates by fixing the sample size (Figures 1-3). Then, we fix the censoring rate to verify the effect of the sample size (Figures 4-6). In conclusion, as one can expect, the quality of fit improves when either the sample size increases or the censoring rate decreases. Also, the quality of estimation seems to be acceptable through all our figures.

4. REFERENCES

[(Chang, 1990)] Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *The Annals of Statistics.* 18(1) :391–404.

[(Chang and Yang, 1987)] Chang, M. N. Yang, G. L. (1987). Strong consistency of a nonpara-

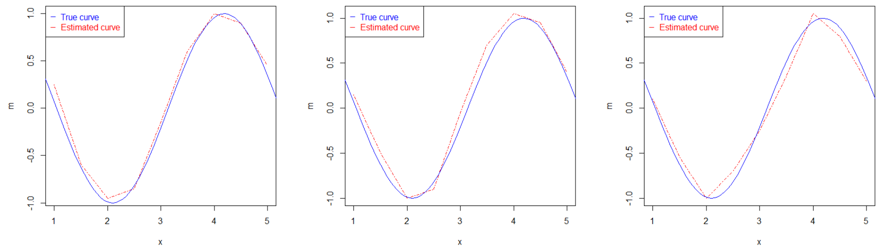

FIGURE 3 – **M3**, $n = 500$ and from left to right CR $= 0\%$, CR $\simeq 30\%$ and CR$\simeq 60\%$

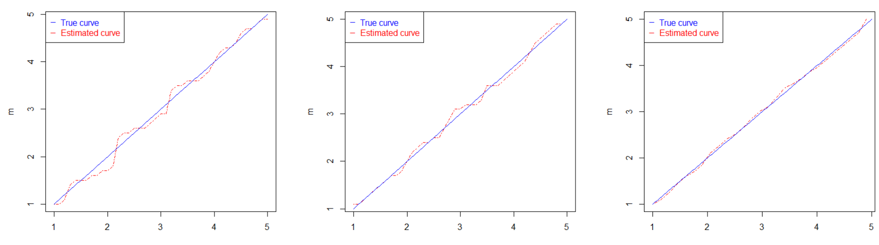

FIGURE 4 – **M1**, CR $\simeq 30\%$ and from left to right $n = 50$, $n = 300$, and $n = 500$

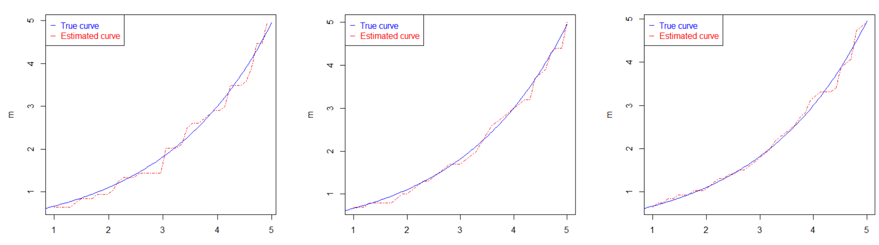

FIGURE 5 – **M2**, CR $\simeq 30\%$ and from left to right $n = 50$, $n = 300$, and $n = 500$

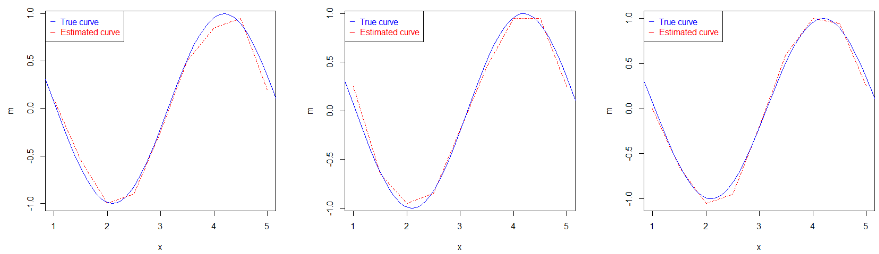

FIGURE 6 – **M3**, CR $\simeq$ 30% and from left to right $n = 300$, $n = 500$, and $n = 800$.

metric estimator of the survival function with doubly censored data. *The Annals of Statistics.* 15(4) :1536–1547.

[(Turnbull, 1974)] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *American Statistical Association.*