EXACT ASYMPTOTIC ERRORS AND BANDWIDTH SELECTION FOR M-ESTIMATION UNDER TRUNCATED-CENSORED AND DEPENDENT DATA

Benseradj Hassiba, Guessoum Zohra**,*

*Faculty of sciences, University of Boumerdes, UMBB;
*Lab MSTD, Faculty of Mathematics, USTHB

ABSTRACT

This work is concerned with the problem of selecting an appropriate bandwidth, for the M-estimator of the robust regression function from left truncated and right censored (LTRC) data, under strong mixing condition. We provide an asymptotic expression for the mean squared error (MSE) of this estimator. As a consequence, a bandwidth selector based on plug-in ideas is introduced. A simulation study is investigated to examine the practical performance of the method.

1. INTRODUCTION

Let $(X_1,Y_1),...,(X_N,Y_N)$ be a sequence of N strongly stationary mixing random variable, identically distributed as (X,Y), taking value in the space $\mathbb{R}^d \times \mathbb{R}$. Our purpose is to study the interaction between X and Y. One may choose, depending the situation under investigation, the conditional mean $\mathbb{E}[Y|X=x]$, which is known to be unstable if outliers are present in the data. Robust regression is involved to overcome this problem, see Huber (1981). More precisely, let ψ_x be a real function which is strictly monotone and integrable on $\mathbb{R}$. For $x \in \mathbb{R}^d$ define the ψ_x-regression function $m_{\psi_x}(.)$ as a zero with respect to θ of

$$\mathbb{E}[\psi_x(Y-\theta)|X=x]=0, \tag{1}$$

In most works where the survival time Y is the variable of interest, referred here as the lifetime, two different problems appear : the first one, if the time origin of the lifetime precedes the start of the study. Only subjects that fail after the beginning of the study are being followed, otherwise they are left truncated. On the other hand, some of these subjects may not be completely observed due to different causes (death for a reason unrelated to the study, or be lost to follow-up), they are then right censored. We are typically in a left truncation and right censoring (LTRC) situation. This type of incomplete data is often encountered in medicine, economics, astronomy,...etc.
We focus in this work, with the problem of selecting a suitable bandwidth needed in kernel M-estimation of the robust regression function, when the response variable Y, is not completely observed, more precisely is subjected to both left truncation and right censorship mechanisms, LTRC model.

2. MODEL AND ESTIMATOR

Let $\{(Y_k,\, T_k,\, W_k), 1 \le k \le N\}$ be a sequence of random vectors from $(Y,\, T,\, W)$, where Y denotes the lifetime under study with continuous distribution function (d.f) F. T and W are the variables of the left truncation and right censoring time with continuous (d.f's) L and G, respectively. Let

$$Z=(Y \wedge W) \quad \text{and} \quad \delta=\mathbf{1}_{\{Y\le W\},}$$

where $(t \wedge u) := \min(t,u)$, and δ is the indicator of censoring status. In random LTRC model one observe $(Z,\ T,\ \delta)$ only if $Z \geq T$. Set $\mu = \mathbb{P}(T \leq Z)$, then we need to assume that $\mu > 0$, otherwise, nothing is observable. Consider the presence of a covariate X, and assume that X admits (d.f) $V(.)$ and a density $v(.)$. Then, denote by $(X_i,\ Z_i,\ T_i,\ \delta_i)$, $i = 1,2,...n$; $(n \leq N)$ a stationary random sample from $(X,\ Z,\ T,\ \delta)$ which one really observe (ie, $T_i \leq Z_i$). Denoted by $\mathbb{P}$ and $\mathbf{P}$ the probability measure related to the N sample, and the actually observed n sample, respectively. Also $\mathbb{E}$ and $\mathbf{E}$ the expectation operators related to $\mathbb{P}$ and $\mathbf{P}$ respectively.
In regression analysis, one expects to identify the relationship between X and Y via the robust regression, this nonparametric model denoted $m(x)$, is implicitly defined as solution of equation (1) where its left hand side can be written as

$$\mathbb{E}\left[\psi_x(Y-\theta)\,|X=x\right] = \frac{\int \psi_x(y-\theta)\, f_{X,Y}(x,y)\, dy}{v(x)},$$

with $v(x) > 0$. Set

$$\Psi_x(x,\theta) := \mathbb{E}\left[\psi_x(Y-\theta)\,|X=x\right] v(x),$$

then $m(.)$ can be viewed as a solution of $\Psi_x(x,\theta) = 0$.
Combining the ideas of robustness with those of smoothed regression, we define a pseudo estimator and a feasible estimator of $\Psi_x(x,\theta)$ respectively by

$$\widetilde{\Psi}_x(x,\theta) := \frac{\mu}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) \frac{\delta_i \psi_x(Z_i-\theta)}{L(Z_i)(1-G(Z_i))},$$

$$\widehat{\Psi}_x(x,\theta) := \frac{\mu_n}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) \frac{\delta_i \psi_x(Z_i-\theta)}{L_n(Z_i)(1-G_n(Z_i))},$$

where $K(.)$ is some kernel function on $\mathbb{R}^d$ and $h_n > 0$ is a bandwidth tending to 0 as $n \to \infty$. G_n, and L_n are the concomitant TJW, see Tsai et al (1987) and the Lynden-Bell, see Lynden (1971) estimators, of the distribution functions G and L respectively. μ_n is a consistent estimate of μ.
Naturally a pseudo estimator and a feasible estimator of $m(x)$ denoted $\widetilde{m}(x)$ and $\widehat{m}(x)$ respectively, are a zero w.r.t θ of

$$\widetilde{\Psi}_x(x,\theta) = 0,\ \ \widehat{\Psi}_x(x,\theta) = 0,$$

respectively. In the sequel, $\{(X_i,\ Z_i,\ T_i,\ \delta_i),\ 1 \leq i \leq n\}$ is assumed to be stationary α-mixing sequences of random vectors. Recall that a sequence $\{\zeta_k, k \geq 1\}$ is said to be α-mixing (strongly mixing) if the mixing coefficient

$$\alpha(n) \overset{def}{:=} \sup_{k\geq 1} \sup \left\{ |\mathbf{P}(AB) - \mathbf{P}(A)\mathbf{P}(B)|\,; A \in F_{n+k}^{\infty}, B \in F_1^k \right\},$$

converge to zero as $n \longrightarrow \infty$ where $F_l^m = \sigma\{\zeta_l, \zeta_{l+1}, ..., \zeta_m\}$ denotes the $\sigma-$algèbra generated by $\zeta_l, \zeta_{l+1}, ..., \zeta_m$ with $l \leq m$.

3. THEORETICAL OPTIMAL BANDWIDTH

A commonly used criterion for selecting a local optimal bandwidth, where h_n is a function of x, $h_n := h_n(x)$, is the mean squared error (MSE) distance, defined by

$$MSE\left(\widetilde{m}(x); h\right) := \mathbf{E}\left[\left(\widetilde{m}(x) - m(x)\right)^2\right],$$

Proposition 1 *Under classical assumptions of the kernel robust regression estimation under α-mixing data, we have*

$$MSE(\widetilde{m}(x)) = Avar + Abias^2 + o\left(h_n^4\right) + o(\frac{1}{nh_n^d})$$

where

$$Avar\left(\widetilde{m}(x)\right) := \frac{1}{nh_n^d} \frac{\Gamma_x\left(x, m\left(x\right)\right)}{\left(\frac{\partial \Psi_x\left(x, m\left(x\right)\right)}{\partial \theta}\right)^2} \int_{\mathbb{R}^d} K^2\left(w\right) dw,$$

$$Abias\left(\widetilde{m}(x)\right) := \frac{h_n^2}{2} \left\{ \frac{-\sum\limits_{1\leq i,j\leq d} \frac{\partial^2 \Psi_x\left(x, m(x)\right)}{\partial x_i \partial x_j} \int\limits_{\mathbb{R}^d} w_i w_j K\left(w\right) dw}{\frac{\partial \Psi_x\left(x, m\left(x\right)\right)}{\partial \theta}} \right\},$$

and $\Gamma_x\left(u, \theta\right) := \mathbb{E}\left[\frac{\mu \psi_x^2\left(Y_1 - \theta\right)}{L\left(Y_1\right) \overline{G}\left(Y_1\right)} | X_1 = u\right] v\left(u\right).$

Theorem 2 *Under the same assumptions as Proposition 1 the optimal local bandwidth which minimize the MSE is given by*

$$h_{MSE}^{opt}\left(x\right) = \frac{1}{n^{1/(d+4)}} \left(\frac{d\Gamma_x\left(x, m\left(x\right)\right) \int\limits_{\mathbb{R}^d} K^2\left(w\right) dw}{\left(\sum\limits_{1\leq i,j\leq d} \frac{\partial^2 \Psi_x\left(x, m(x)\right)}{\partial x_i \partial x_j} \int\limits_{\mathbb{R}^d} w_i w_j K\left(w\right) dw \right)^2} \right)^{\frac{1}{d+4}}. \tag{2}$$

4. ITERATIVE PLUG-IN BANDWIDTH SELECTION

Based on the idea of "plugging-in" appropriate estimators of the unknown quantities, that appear in the formula of the *AMSE* - optimal bandwidth $\left(h_{AMSE}^{opt}\right)$ given in formula (2) is estimated (with gaussian kernel, $d = 1$) by

$$\widehat{h}_{Plug}^{opt} := \left(\frac{\widehat{\Gamma}_x\left(x, \widehat{m}\left(x\right)\right)}{2n\sqrt{\pi} \left(\frac{\partial^2 \widehat{\Psi}_x\left(x, \widehat{m}(x)\right)}{\partial x^2} \right)^2} \right)^{1/5} \tag{3}$$

where

$$\widehat{\Gamma}_x\left(x, \theta\right) := \frac{\mu_n}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \frac{\delta_i \psi_x^2\left(Z_i - \theta\right)}{L_n^2\left(Z_i\right) \overline{G_n}^2\left(Z_i\right)},$$

and

$$\frac{\partial^2 \widehat{\Psi}_x\left(x, \theta\right)}{\partial x^2} = \frac{\mu_n}{nh_n} \sum_{i=1}^{n} K''\left(\frac{x - X_i}{h_n}\right) \frac{\delta_i \psi_x\left(Z_i - \theta\right)}{L_n\left(Z_i\right) \overline{G}_n\left(Z_i\right)},$$

$K''\left(.\right)$ is the second derivative of the kernel $K\left(.\right)$.
The "Iterative" approach, proposed first by Gasser et al (1991) for the classical regression in the complete case, adapted to the Robust regression context from LTRC data (a little different), is based on the iteration algorithm.

Iterative Plug-in Algorithm

step 1 Initialize $h_n^{(0)} \asymp n^{-1/5}$, since this is the bandwidth which is candidate for being the optimal global bandwidth.

step 2 Iterate i=1,2,..., and estimate $\widehat{m}^{(i)}(x)$, using $h_n^{(i-1)}$, for each fixed x.

step 3 Estimate $\widehat{\Gamma}_x^{(i)}$ and $(\frac{\partial^2 \widehat{\Psi}_x}{\partial x^2})^{(i)}$, using $h_n^{(i-1)}$.

step 4 Find $h_n^{(i)}$ using expression (3) with the estimators calculated in step 2 and 3.

step 5 Stop criterion : $\left|h_n^{(i-1)} - h_n^{(i)}\right| < \varepsilon$, and set $\widehat{h}_{Plug}^{opt} = h_n^{(i)}$, where $\varepsilon > 0$ is a precision needed.

5. CONCLUSIONS

The plug-in method builds on estimating the asymptotically optimal bandwidth from the data. Since estimators for the residual variance and for an asymptotic expression for the bias are plugged into the asymptotic formula, such selection rules are called "plug-in" estimators. The iterative plug-in method is fast, flexible and present a good performance in simulation.

References

Gasser, T., Kneip, A. and Köhler, W. (1991). A fexible and fast method for automatic smoothing. J. Amer. Statist. Assoc, 86, 643-652.

Huber, P.J. (1981) Robust Statistics. John Wiley & Sons, New York.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3cr quasars. Monthly Notices Roy. Astronom Soc., 155, 95-118

Tsai, W. Y, Jewell, N. P.and Wang, M. C. (1987). A note on the product limit estimator under right censoring and left truncation. Biometrika, 74, 883-886.