OPTIMAL BANDWIDTH SELECTION IN M-TYPE ESTIMATE OF THE REGRESSION FUNCTION IN ASSOCIATED AND LEFT-TRUNCATED MODEL

Gheliem Asma,Guessoum Zohra***

*M.S.T.D Laboratory, USTHB , Algiers, Algeria,
agheliem@usthb.dz
**M.S.T.D Laboratory, USTHB , Algiers, Algeria,
zguessoum@usthb.dz

ABSTRACT

The choice of the smoothing parameter, or bandwidth, is crucial to the effective performance of the estimator. In this contribution we are interested by a bandwidth-selection rule in the M-type estimation of the regression function in associated and left-truncated model .

1. INTRODUCTION AND MOTIVATION

Let Y be a real random variable (rv) of interest with distribution function (df) F and X a random vector of covariates taking its values in $\mathbb{R}^d$ with (df) V and continuous density v and we want to estimate Y after observing X. The regression function between Y and X for $x \in \mathbb{R}^d$, is defined by the conditional expectation of Y given $X = x$, that is

$$r(x) = \mathbb{E}[Y|X = x].$$

Note that the function $r(x)$ can be expressed as

$$r(x) = \arg\min_{s\in\mathbb{R}} \mathbb{E}[(Y_i - s)^2|X = x].$$

This latter is a particular case of a more general definition when dealing with robust estimation, viz

$$r(x) = \arg\min_{s\in\mathbb{R}} \mathbb{E}[\rho(Y_i - s)|X = x],$$

where $\rho(.)$ is an outlier-resistant loss and convex function defined on $\mathbb{R}$, hence, one can see r(x) as a solution, with respect to s (w.r.t.s), of

$$\mathbb{E}[\psi(Y_i - s)|X = x] = 0$$

where $\psi(.) := \frac{\partial}{\partial s}\rho(.)$ is a monotone (score) function.

The corresponding non parametric M-estimator is equivalent to solving the equation

$$\sum_{i=1}^{n} K_d\left(\frac{x - X_i}{h_n}\right)\psi(Y_i - s) = 0,$$

where K_d is a kernel function on $\mathbb{R}^d$ and h_n is a sequence of positive real numbers which goes to zero as n goes to infinity (bandwidth). As it is well known, this last estimator is sensitive to the

presence of outliers (for example in economic and finance time series) and suffers from being not robust.
In realistic framework, the variable of interest Y may be subject to censoring and/or truncation. Under random left-truncation model (RLTM), Wang and Liang (2012) [3] constructed the M-estimator of the non parametric regression function for α-mixing data and truncated multivariate data and established a weak and strong consistency of the estimator (without rate) as well as its asymptotic normality.

The choice of the smoothing parameter, or bandwidth, is crucial to the effective performance of the estimator. In the complete sample case, several authors have taken an interest to the mean integrated squared error (MISE) which has an asymptotic decomposition as a simple variance term, a simple squared bias tern and some negligible terms.

Our focus in this contribution is to see how this type of decomposition may be done for an M-type regression estimator, in the case of truncated and associated data and to give the theoretical form (local and global) of the bandwidth.
The concept of association was introduced and defined by Esary 1967[1],

Definition 1 *A set of finite family of rv's $(X_1, X_2, ..., X_N)$ is said to be associated if for every pair of functions $g_1(.)$ and $g_2(.)$ from $\mathbb{R}^N$ to $\mathbb{R}$, which are non decreasing componentwise,*

$$Cov(g_1(X), g_2(X)) \geq 0,$$

whenever the covariance is defined, where $X = (X_1, X_2, ..., X_N)$. An infinite sequence $\{X_N, N \geq 1\}$ of rv's is said to be associated if every finite subset is associated.

2. MODEL AND MAIN RESULT

Let $(X_k, Y_k), 1 \leq k \leq N$ be a sequence of associated random vector, where Y has continuous df F and T be the truncation variable with continuous df G, defined on the same probability space $(\Omega, F, \mathbb{P})$. Let f (.,.) be the joint density function of the random vector (X, Y). We assume throughout this paper that T and (X, Y) are independent.
Under RLTM, the lifetime Y and T are observable only when $Y \geq T$, here N is the potential sample size. As a consequence of truncation, the size N is fixed but unknown and n, the size of the actually observed sample, is random and known with $n \leq N$.

Let $\mu =: \mathbf{P}(Y \geq T)$ be the probability to observe the rv of interest Y. Under RLTM, we denote by m(x) the implicit solution (w.r.t.s), of

$$H(x,s) := \int_{\mathbb{R}} \psi(y-s) \frac{\mathbf{f}(x,y)}{G(y)} dy = \frac{1}{\mu} \mathbb{E}[\psi(Y-s)|X=x)] v(x) = 0.$$

Moreover as $H(x,s)$ can be empirically estimated by

$$\tilde{H_n}(x,s) := \frac{1}{nh_n^d} \sum_{i=1}^{n} \frac{1}{G(Y_i)} K_d \left(\frac{x - X_i}{h_n} \right) \psi(Y_i - s),$$

we propose $\tilde{m}_n(x)$, the implicit solution (w.r.t.s) of $\tilde{H_n}(x,s) = 0$, as un M-estimator of $m(x)$. Nevertheless, as G(.) is unknown, the estimator $\tilde{H}_n(x,s)$ is unusable and so is $\tilde{m}_n(x)$. We finally define $\hat{m}_n(x)$, the implicit solution (w.r.t.s) of

$$\hat{H}_n(x,s) := \frac{1}{nh_n^d} \sum_{i=1}^{n} \frac{1}{G_n(Y_i)} K_d \left(\frac{x - X_i}{h_n} \right) \psi(Y_i - s) = 0,$$

as a feasible M-estimator of $m(x)$, where $G_n(x)$ is the well known product limit estimator of $G(x)$ in RLTM, proposed by Lynden-Bell(1971)[2] defined by

$$G_n(y) \quad = \quad \prod_{T_i > y} \left[\frac{nC_n(T_i) - 1}{nC_n(T_i)} \right],$$

where

$$C_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{T_i \le y \le Y_i\}},$$

is an estimator of $C(y) := \mathbb{P}\{T \le y \le Y | Y \ge T\}$. Set
$\Lambda_1(x,u,m(x)) := \frac{1}{\mu}\mathbb{E}\Big[\psi(Y - m(x))\Big| X = u\Big] v(u) = H(x,m(x)) = 0,$
$\Lambda_2(x,u,m(x)) := \frac{1}{\mu}\mathbb{E}\Big[\psi^2(Y - m(x))G^{-1}(Y)|X = u\Big] v(u),$
$\Lambda_3(x,u,m(x)) := \frac{1}{\mu}\mathbb{E}\Big[\psi'(Y - m^*(x))|X = u\Big] v(u),$
$\Lambda_4(x,u,m(x)) := \frac{1}{\mu}\mathbb{E}\Big[\psi'^2(Y - m^*(x))G^{-1}(Y)|X = u\Big] v(u)$

To simplify the notations, let
$\Lambda_i := \Lambda_i(x,x,m(x)) \quad i = 1,2,3,4$
Then we have,

Proposition 1 *Under some regularity conditions, we have*

$$\mathbf{Var}(\tilde{m}_n(x)) \quad = \quad \frac{\kappa}{nh_n^d}\frac{\Lambda_2}{\Lambda_3^2} + o\left(\frac{1}{nh_n^d}\right)$$

With $\kappa := \int_{\mathbb{R}^d} K_d^2(t)dt (< +\infty)$

Proposition 2 *Under some regularity conditions, we have*

$$Bias(\tilde{m}_n(x)) \quad = \quad -\frac{\frac{h_n^2}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2\Lambda_1}{\partial x_i \partial x_j}\int_{\mathbb{R}^d} t_i^2 K_d(t)dt}{\Lambda_3} + 0\left(\frac{1}{nh_n^d}\right)$$

Theorem 3 **i)** *Under the conditions of Propositions 1 and 2, we get*

$$a)\ MSE(x,h_n) \quad = \quad \frac{h_n^4}{4}\frac{\left(\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2\Lambda_1}{\partial x_i \partial x_j}\int_{\mathbb{R}^d} t_i^2 K_d(t)dt\right)^2}{\Lambda_3^2}$$
$$+ \quad \frac{\kappa}{nh_n^d}\left(\frac{\Lambda_2}{\Lambda_3^2}\right) + o\left(\frac{1}{nh_n^d}\right)$$

$$b)\ AMSE(x,h_n) \quad = \quad \frac{h_n^4}{4}\frac{\left(\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2\Lambda_1}{\partial x_i \partial x_j}\int_{\mathbb{R}^d} t_i^2 K_d(t)dt\right)^2}{\Lambda_3^2}$$
$$+ \quad \frac{\kappa}{nh_n^d}\left(\frac{\Lambda_2}{\Lambda_3^2}\right)$$

ii) *Under the conditions of Propositions 1 and 2 and for $I \subseteq \mathbb{R}$, we get*

$$MISE(x,h_n) \quad = \quad \int_I \left\{ \frac{h_n^4}{4} \frac{\left(\sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 \Lambda_1}{\partial x_i \partial x_j} \int_{\mathbb{R}^d} t_i^2 K_d(t)dt \right)^2}{(\Lambda_3)^2} + \frac{\kappa}{nh_n^d} \left(\frac{\Lambda_2}{\Lambda_3^2} \right) \right\} dx + o\left(\frac{1}{nh_n^d} \right)$$

Corollary 4 **i)** *Under the conditions of Propositions 1 and 2, we get*

$$h_{n,MSE}^{opt} = n^{\frac{-1}{d+4}} \cdot \left[\frac{d\kappa\Lambda_2}{\left(\sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 \Lambda_1}{\partial x_i \partial x_j} \int_{\mathbb{R}^d} t_i^2 K_d(t)dt \right)^2} \right]^{\frac{1}{d+4}}$$

ii) *Under some regularity conditions*

$$h_{n,MISE}^{opt} = n^{\frac{-1}{d+4}} \cdot \left[\frac{d \int_I \kappa\Lambda_2 dx}{\int_I \left(\sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 \Lambda_1}{\partial x_i \partial x_j} \int_{\mathbb{R}^d} t_i^2 K_d(t)dt \right)^2 dx} \right]^{\frac{1}{d+4}}$$

3. REFERENCES

[1] Esary(1967). Association of random variables with applications. *Ann. Math. Stat.* Vol **38** 1466–1476.

[2] LYNDEN-BELL D. (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices Royal Astronomy Society* **155** 95–118.

[3] WANG, J.F. and LIANG, H.Y. (2012). Asymptotic properties for an M-estimator of the regression function with truncation and dependent data. *J. Korean Stat. Soc.* **41** 351–367.