Journal of Multivariate Analysis 102 (2011) 448–467

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# Wavelet estimation of conditional density with truncated, censored and dependent data

Han-Ying Liang [a,b,*], Jacobo de Uña-Álvarez [b]

[a] Department of Mathematics, Tongji University, Shanghai 200092, PR China

[b] Department of Statistics and OR, Facultad de Ciencias Econmicas y Empresariales, Universidad de Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Spain

ARTICLE INFO

*Article history:*
Received 25 February 2010
Available online 16 October 2010

*AMS 2010 subject classifications:*
62G07
62G20

*Keywords:*
Mean integrated squared error
Asymptotic normality
Nonlinear wavelet estimator
Conditional density
Truncated and censored data
$\alpha$-mixing

ABSTRACT

In this paper we define a new nonlinear wavelet-based estimator of conditional density function for a random left truncation and right censoring model. We provide an asymptotic expression for the mean integrated squared error (MISE) of the estimator. It is assumed that the lifetime observations form a stationary $\alpha$-mixing sequence. Unlike for kernel estimators, the MISE expression of the wavelet-based estimators is not affected by the presence of discontinuities in the curves. Also, asymptotic normality of the estimator is established.



## 1. Introduction

In recent years, wavelet methods in nonparametric curve estimation have become a well-known and powerful technique. We refer to the monograph by Härdle et al. [18] for a systematic discussion of wavelets and their applications in statistics. The major advantage of the wavelet method is its adaptability (in the minimax sense) to the degree of smoothness of the underlying unknown curve. These wavelet estimators typically achieve the optimal convergence rates over exceptionally large function spaces. For more information and related references, see the initial works by Kerkyacharian and Picard [20,21], Donoho and Johnstone [6,7], and Donoho et al. [8,9]. Hall and Patil [17] gave for the first time an asymptotic expression of the mean integrated squared error (MISE) of a nonlinear wavelet density estimator, and compared its performance to that corresponding to the kernel density estimator. These authors showed that the asymptotic MISE formula is the same in both the smooth and nonsmooth density cases, a fact that is not true for the kernel method.

In medical follow-up or in engineering life testing studies, one may not be able to observe the variable of interest, referred to hereafter as the lifetime. Among the different forms in which incomplete data appear, right censoring and left-truncation are two common ones. Some authors have studied wavelet density estimation with censored data. For example, Antoniadis et al. [2] considered linear wavelet density estimation under random censoring, and provided an asymptotic MISE convergence rate under smoothness assumptions on the underlying density function. Li [23] proposed a nonlinear wavelet density estimator with censored data and derived a result similar to the main result, Theorem 2.1, of Hall and Patil [17], for the MISE; see also [28] who considered the Koziol–Green model of random censorship. All of the above works are devoted

* Corresponding author at: Department of Mathematics, Tongji University, Shanghai 200092, PR China.
*E-mail addresses:* hyliang83@yahoo.com (H.-Y. Liang), jacobo@uvigo.es (J. de Uña-Álvarez).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.10.004

to the independent setting. For the dependent case, Liang et al. [24] discussed the global $L_2$ error of the nonlinear wavelet estimators of the density function in the Besov space under censoring and stationary $\alpha$-mixing assumptions; for complete data, Truoug and Patil [30] gave an asymptotic expression of the MISE in nonlinear wavelet regression with $\alpha$-mixing data. However, the construction and asymptotic properties of the nonlinear wavelet estimator of the conditional density function for the left truncated and right censored (LTRC) model are not available in the literature so far. Moreover, we are unaware of any paper dealing with wavelet estimation of a conditional density in the simplest situation of i.i.d. complete data.

Let $(Y, T, W)$ be a random vector, where $Y$ is the lifetime with distribution function (df) $F$, $T$ is the random left truncation time with the df $L$ and $W$ denotes the random right censoring time with the df $G$. In the random LTRC model one observes $(Z, T, \delta)$ if $Z \geq T$, where $Z = \min(Y, W)$ and $\delta = I(Y \leq W)$. When $Z < T$ nothing is observed. Clearly, if $Y$ is independent of $W$, then $Z$ has the df $H = 1 - (1 - F)(1 - G)$. Take $\theta = P(T \leq Z)$, then necessarily, we assume $\theta > 0$. Let $(Z_i, T_i, \delta_i)$, for $i = 1, 2, \ldots, n$, be a stationary random sample from $(Z, T, \delta)$ which one observes then $(T_i \leq Z_i, \forall i)$. The product-limit estimator (PLE), $F_n$, of $F$ is defined in [31] as follows:

$$1 - F_n(y) = \prod_{i=1}^{n} \left(1 - \frac{I(Z_i \leq y, \delta_i = 1)}{nC_n(Z_i)}\right), \tag{1.1}$$

where $C_n(y) = n^{-1} \sum_{i=1}^{n} I(T_i \leq y \leq Z_i)$ is the empirical estimator of $C(y) = P(T \leq y \leq Z | T \leq Z)$. Under independent and identically distributed (i.i.d.) assumptions, the properties of $F_n$ have been studied by Wang [34], Gu and Lai [15], Lai and Ying [22], Gijbels and Wang [11], and Zhou [36], among others. Nonparametric estimates of the density and hazard rate for $F$ have been studied by Gijbels and Wang [11], Sun and Zhou [29], Uzunoğullari and Wang [32], Gu [14] and Zhou et al. [37].

Let $\mathbf{X}$ be a $\mathbb{R}^d$-valued random vector of covariates related with $Y$. Assume that $\mathbf{X}$ admits df $M(\cdot)$ and density $m(\cdot)$. Then denote by $(\mathbf{X}_i, Z_i, T_i, \delta_i)$, $i = 1, 2, \ldots, n$ a stationary random sample from $(\mathbf{X}, Z, T, \delta)$ which one observes $(T_i \leq Z_i, \forall i)$. Without loss of generality, we assume that $Y$, $T$ and $W$ are nonnegative random variables, as usual in survival analysis. Following the idea of Iglesias-Pérez and González-Manteiga [19], we define a generalized product-limit estimator, GPLE, of the conditional survival function of $Y$ given $\mathbf{X} = \mathbf{x}$ for the LTRC data, given by

$$1 - \hat{F}_n(y|\mathbf{x}) = \prod_{i=1}^{n} \left(1 - \frac{I(Z_i \leq y)\delta_i B_{ni}(\mathbf{x})}{\sum_{j=1}^{n} I(T_j \leq Z_i \leq Z_j) B_{nj}(\mathbf{x})}\right), \tag{1.2}$$

where $B_{ni}(\mathbf{x}) = K(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}) / \sum_{j=1}^{n} K(\frac{\mathbf{x}-\mathbf{X}_j}{h_n})$, $K$ denotes a kernel function on $\mathbb{R}^d$, and $0 < h_n \to 0$ is the bandwidth parameter. Note that the GPLE reduces to the estimator for left truncated data when there is no right censoring ($\delta = 1, Z = Y$) (see [1]), and to the estimator for right censored data when there is no left truncation ($T = 0$) (see [3,12]). On the other hand, in the absence of covariates, the GPLE reduces to the PLE estimator defined by (1.1) when $(B_{ni}(\mathbf{x}) = 1/n, \forall i)$. Under i.i.d. assumptions and for the case of $d = 1$, Iglesias-Pérez and González-Manteiga [19] obtained an almost sure representation and asymptotic normality of $\hat{F}_n(y|\mathbf{x})$. Asymptotic results for this estimator under mixing conditions were stated in [25].

Let the conditional df of $Y$ given $\mathbf{X} = \mathbf{x}$ be $F(y|\mathbf{x}) = P(Y \leq y | \mathbf{X} = \mathbf{x})$, and its density function be $f(y|\mathbf{x})$. In view of (1.2), we, in this paper, define a new nonlinear wavelet-based estimator of $f(y|\mathbf{x})$ for the LTRC model, and establish the MISE as well as the asymptotic normality of the estimator with dependent data.

Next, for any df $Q(y) = P(\eta \leq y)$, its density function is denoted by $q(y)$. We denote the left and right support endpoints by $a_Q = \inf\{y : Q(y) > 0\}$ and $b_Q = \sup\{y : Q(y) < 1\}$, respectively. Define

$$H_1(y|\mathbf{x}) = P(Z \leq y, \delta = 1 | \mathbf{X} = \mathbf{x}), \qquad \theta(\mathbf{x}) = P(T \leq Z | \mathbf{X} = \mathbf{x}),$$
$$C(y|\mathbf{x}) = P(T \leq y \leq Z | \mathbf{X} = \mathbf{x}, T \leq Z), \qquad \Lambda(y|\mathbf{x}) = \int_0^y \frac{\mathrm{d}F(t|\mathbf{x})}{1 - F(t^-|\mathbf{x})} = -\ln(1 - F(y^-|\mathbf{x})).$$

Also, we define $Q(y|\mathbf{x}) = P(\eta \leq y | \mathbf{X} = \mathbf{x})$ and $Q^*(y) = P(\eta \leq y | T \leq Z)$, with density functions denoted by $q(y|\mathbf{x})$ and $q^*(y)$, respectively. Then we have $M^*(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x} | T \leq Z)$, $H_1^*(y|\mathbf{x}) = P(Z \leq y, \delta = 1 | \mathbf{X} = \mathbf{x}, T \leq Z)$.

**Remark 1.1.** It is easy to verify that $m^*(\mathbf{x}) = \theta^{-1}\theta(\mathbf{x})m(\mathbf{x})$. Assume that $Y$, $T$ and $W$ are conditionally independent given $\mathbf{X} = \mathbf{x}$, then $\Lambda(y|\mathbf{x}) = \int_0^y \frac{\mathrm{d}H_1^*(t|\mathbf{x})}{C(t|\mathbf{x})}$ for $y < b_{H(\cdot|\mathbf{x})}$, and

$$C(y|\mathbf{x}) = \theta^{-1}(\mathbf{x})L(y|\mathbf{x})(1 - G(y|\mathbf{x}))(1 - F(y|\mathbf{x})), \qquad H_1^*(y|\mathbf{x}) = \theta^{-1}(\mathbf{x}) \int_0^y L(t|\mathbf{x})(1 - G(t|\mathbf{x}))f(t|\mathbf{x})\mathrm{d}t.$$

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

$$\alpha(n) := \sup_{k \geq 1} \sup\{|P(AB) - P(A)P(B)| : A \in \mathcal{F}_{n+k}^{\infty}, B \in \mathcal{F}_1^k\}$$