Statistics and Probability Letters 96 (2015) 255–261

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Rate of strong consistency for nonparametric estimators based on twice censored data

Abderrahim Kitouni *, Mohamed Boukeloua, Fatiha Messaci

Département de Mathématiques, Université Constantine 1, route d'Ain El Bey, 25017 Constantine, Algeria

ARTICLE INFO

Article history:
Received 12 May 2014
Received in revised form 20 August 2014
Accepted 5 October 2014
Available online 12 October 2014

Keywords:
Almost complete convergence
Product-limit estimator
Density
Failure rate
Twice censoring

ABSTRACT

We establish the uniform almost complete convergence, with rate, for estimators of both the cumulative hazard and the survival functions when the data are subject to twice censoring. Then, we derive similar convergence results for kernel type estimators of the density and the failure rate.

1. Introduction

Many works in the statistical literature deal with nonparametric estimation when the variable of interest is either complete or singly censored. However, in reliability and survival time studies, one can encounter a more complicated random censorship situation. Such is the case when the variable of interest X is right censored by a variable R, $\min(X, R)$ is left censored by a variable L, and X, R and L are independent.

An example of such a model, given in Patilea and Rolin (2006), is to consider a reliability system consisting of three components C_1, C_2 and C_3 with C_1 and C_2 in series and C_3 in parallel with the series system. The lifetimes of the components C_1, C_2 and C_3, respectively denoted by X, R and L are assumed to be independent and we can determine which component failed at the same time as the system. So, instead of observing X, we can only observe $\max(\min(X, R), L)$ and a censoring indicator.

In this situation, one way to estimate the distribution function of the lifetime of interest X is by means of the product-limit estimator F_n defined by Patilea and Rolin (2006) who established its uniform almost sure convergence under some identifiability conditions. Messaci and Nemouchi (2011, 2013) precised the rate of this convergence, which is of order $\sqrt{\log\log n}/\sqrt{n}$. In this work, we show the uniform almost complete convergence, with rate $O(\sqrt{\log n}/\sqrt{n})$, for the estimator F_n under the same conditions as Messaci and Nemouchi (2013) but the hypothesis of the continuity of the distribution functions of the latent variables has been relaxed. Notice that the almost complete convergence is stronger than the almost sure one. In other words, we improve the convergence mode and we dispense with the hypothesis of continuity, at the cost of deteriorating the rate of convergence. We also give the same rate of convergence for an estimator of the cumulative hazard function. Moreover our present results are applied to derive the uniform almost complete convergence for estimators of both the density and the failure rate of the twice censored variable X. Our approach consists of bringing back the above twice censorship model to both the complete and the single censorship ones in order to apply known results in these cases.

* Corresponding author. Tel.: +213 31 81 90 08; fax: +213 31 81 90 11.
E-mail addresses: a.kitouni@gmail.com (A. Kitouni), boukeloua.mohamed@gmail.com (M. Boukeloua), f_messaci@yahoo.fr (F. Messaci).

http://dx.doi.org/10.1016/j.spl.2014.10.005
0167-7152/

This is the subject of Section 2. Section 3 is reserved to the application of our previous result to establish rates of uniform almost complete convergence for kernel estimators of both the density and the failure rate of the twice censored variable X.

Throughout this paper the following notations will be adopted. Let $F_V(t) = P(V \leq t)$ and $S_V(t) = 1 - F_V(t)$ denote respectively the distribution and the survival functions of a real random variable (r.r.v.) V, and let $T_V = \sup\{t \in \mathbb{R}/F_V(t) < 1\}$ and $I_V = \inf\{t \in \mathbb{R}/F_V(t) > 0\}$ denote respectively the upper and the lower endpoints of the support of F_V. Furthermore, for any right continuous function $\varphi : \mathbb{R} \to \mathbb{R}$, we set $\varphi(t^-) = \lim_{\varepsilon \overset{>}{\to} 0} \varphi(t - \varepsilon)$ and $\Delta\varphi(t) = \varphi(t) - \varphi(t^-)$ whenever the limit exists.

For easy reference, recall the following definitions. Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of r.r.v. We say that $(Z_n)_{n\in\mathbb{N}}$ converges almost completely (a.co.) to some r.r.v. Z if $\forall\varepsilon > 0$, $\sum_{n=1}^{\infty} P(|Z_n - Z| > \varepsilon) < \infty$. Moreover, let $(u_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers going to zero; we say that the rate of the almost complete convergence of $(Z_n)_{n\in\mathbb{N}}$ to Z is of order (u_n) and we note $Z_n - Z = O_{\text{a.co.}}(u_n)$ if $\exists\, \varepsilon > 0$, $\sum_{n=1}^{\infty} P(|Z_n - Z| > \varepsilon u_n) < \infty$. This notion and its properties are discussed at length in Ferraty and Vieu (2006). Remark that one can straightforwardly derive, from Borel–Cantelli lemma, that this convergence is stronger than the almost sure one.

2. Estimation of the distribution function

We first start by recalling the product-limit estimator of the distribution function of a left censored variable. To this aim, let L and Y be independent positive random variables representing respectively the lifetime of interest and a left censoring variable. In this setting, recall that instead of observing L, one can only have at disposal a sample $(Z_i, \delta_i)_{1\leq i\leq n}$ of independent random copies of (Z, δ) where $Z = \max(L, Y)$ and $\delta = 1_{\{L\geq Y\}}$ (where $1_{\{.\}}$ denotes the indicator function). Denote by H the distribution function of Z and by N its subdistribution function defined by $N(t) = P(Z \leq t, \delta = 1)$. Consider the empirical versions of H and N given respectively by

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i\leq t\}} \quad \text{and} \quad N_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i\leq t, \delta_i=1\}}. \tag{1}$$

The relation $F_L(t) = \prod_{x>t}(1-d\Gamma(x))$, where $\prod$ is the product integral (see Gill and Johansen, 1990) and $\Gamma(t) = -\int_t^{\infty} \frac{dF_L}{F_L} = -\int_t^{\infty} \frac{dN}{H}$, suggests to estimate F_L by

$$\widetilde{F}_n(t) = \prod_{j/Z'_j>t} (1 - \Delta\Gamma_n(Z'_j)), \tag{2}$$

where $\Gamma_n(t) = -\int_t^{\infty} \frac{dN_n}{H_n}$ and $\{Z'_j,\ 1 \leq j \leq m\}$ are the distinct values in increasing order of $\{Z_i,\ 1 \leq i \leq n\}$.

Note that the same estimator can be rediscovered from the Kaplan–Meier one by reversing time, so we can adapt Theorem 1 of Bitouzé et al. (1999) to get

$$P\left(\sup_{I_Y<t} |\widetilde{F}_n(t) - F_L(t)| > \sqrt{\frac{\log n}{n}}\right) \leq 2.5\, e^{-2\log n + C\sqrt{\log n}},$$

where C is an absolute constant. Choosing $\alpha < 1$, for n large enough we obtain $e^{C\sqrt{\log n}} \leq n^{\alpha}$, which implies that

$$\sup_{I_Y<t} |\widetilde{F}_n(t) - F_L(t)| = O_{\text{a.co.}}\left(\sqrt{\frac{\log n}{n}}\right). \tag{3}$$

It is worth noticing that

$$\sup_{t\in\mathbb{R}} |H_n(t) - H(t)| = O_{\text{a.co.}}\left(\sqrt{\frac{\log n}{n}}\right), \tag{4}$$

in view of the DKW inequality (Dvoretzky et al., 1956; Massart, 1990).

Now, let X, R and L be independent positive variables representing respectively the lifetime of interest, a right censoring variable, and a left censoring one operating on $\min(X, R)$. We are in the situation where one can only observe a sample of $(\max(\min(X_i, R_i), L_i), A_i)$ of n independent and identically distributed variables of the pair $(\max(\min(X, R), L), A)$ where $A = 1_{\{L<R<X\}} + 2 \times 1_{\{\min(X,R)\leq L\}}$. This is the twice censoring model studied in Patilea and Rolin (2006).

Set $Y = \min(X, R)$ and $Z = \max(Y, L) = \max(\min(X, R), L)$, and define the subdistribution function of H for uncensored observations by $H^{(0)}(t) = P(Z \leq t, A = 0)$. Since $H^{(0)}(t) = \int_0^t F_L(x^-)S_R(x^-)\, dF_X(x)$ and $H(t) = F_L(t)F_Y(t)$, the cumulative hazard function Λ of X can be written for any t such that $I_L < t < T_R$ as

$$\Lambda(t) = \int_0^t \frac{dF_X(u)}{S_X(u^-)} = \int_0^t \frac{dH^{(0)}(u)}{F_L(u^-) - H(u^-)}.$$

Let $H_n^{(0)}$ be the empirical version of $H^{(0)}$ given by

$$H_n^{(0)}(t) = \frac{1}{n}\sum_{1\leq i\leq n} 1_{\{Z_i\leq t, A_i=0\}}.$$