Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
C. R. Acad. Sci. Paris, Ser. I 337 (2003) 207–212
COMPTES RENDUS MATHEMATIQUE

Probability Theory/Statistics

Chung–Smirnov property for smoothed distribution function estimator under random censorship

Elias Ould-Saïd, Ouafae Yazourh-Benrabah

Université du Littoral Côte d'Opale, LMPA, centre de la Mi-Voix, BP 699, 62228 Calais, France

Received 10 January 2003; accepted after revision 17 June 2003

Presented by Paul Deheuvels

Abstract

Let $(X_n)_{n\geqslant 1}$ be a sequence of independent and identically distributed (iid) random variables (rv) with common distribution function (df) F and another iid sequence $(C_n)_{n\geqslant 1}$ with df G independent of $(X_n)_{n\geqslant 1}$. Here we consider the Smoothed Kaplan–Meier Estimator $\dot{F}_n$ of F defined as integral of nonparametric density estimators. It is shown that if F satisfies some smoothness conditions, $\dot{F}_n$ has the Chung–Smirnov property, that is, with probability one,

$$\limsup_{n\to\infty}\left(\frac{2n}{\log\log n}\right)^{1/2}\|\dot{F}_n - F\|_T = C_{F,G},$$

where $C_{F,G}$ is a constant depending only on F and G ($\|\cdot\|_T$ and T are defined below). In this Note, we extend the result of Winter (1979) and Degenhardt (1993) to the censorship model and those of Csörgö and Horvath (1983) to the smoothed estimator with the same constant $C_{F,G}$. ***To cite this article: E. Ould-Saïd, O. Yazourh-Benrabah, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

Résumé

Propriété de Chung–Smirnov de l'estimateur lissé de la fonction de répartition en présence de censure. On considère une suite de variables aléatoires iid $(X_n)_{n\geqslant 1}$ de même fonction de répartition (f.d.r.) F et une autre suite de variables aléatoires $(C_n)_{n\geqslant 1}$ de f.d.r. G indépendantes de $(X_n)_{n\geqslant 1}$. On considère un estimateur lissé par convolution $\dot{F}_n$ de F. Nous montrons que cet estimateur vérifie la propriété de Chung–Smirnov. Dans cette Note, nous étendons les résultats de Winter (1979) et Degenhardt (1993) au cas censuré et celui de Csörgö et Horvath (1983) à l'estimateur lissé avec la même constante $C_{F,G}$. ***Pour citer cet article : E. Ould-Saïd, O. Yazourh-Benrabah, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

Version française abrégée

Dans les modèles de durée, on dispose de deux suites $(X_n)_{n\geqslant 1}$ et $(C_n)_{n\geqslant 1}$ mutuellements indépendantes de f.d.r. F et G respectivement. Les v.a. observables sont $Z_i = \min\{X_i, C_i\}$ et $\delta_i = \mathbb{1}_{\{X_i\leqslant C_i\}}$ pour $i = 1, 2, \ldots, n$. Il

E-mail addresses: ouldsaid@lmpa.univ-littoral.fr (E. Ould-Saïd), benrabah@lmpa.univ-littoral.fr (O. Yazourh-Benrabah).

1631-073X/$ – see front matter
doi:10.1016/S1631-073X(03)00311-X

est bien connu que pour ce type de modèle, l'estimateur de la f.d.r. F est l'estimateur de Kaplan–Meier [2] $\widehat{F}_n$, défini ci-dessous (3). Si l'on suppose que F est continue, il est naturel de considérer un estimateur lissé de F au lieu de l'estimateur en escalier $\widehat{F}_n$. De tels estimateurs apparaissent naturellement comme des intégrales d'estimateurs nonparamétriques de la densité, définis par

$$\hat{f}_n(t) = \frac{1}{a_n} \int K\left(\frac{t-s}{a_n}\right) \mathrm{d}\widehat{F}_n(s), \tag{1}$$

où K est une densité de probabilité telle que $K \geqslant 0$, $\int K(u)\,\mathrm{d}u = 1$, et a_n est une suite de réels positifs décroissant vers zéro quand n tend vers l'infini et où $\widehat{F}_n$ est l'estimateur de Kaplan–Meier défini en (3).

L'estimateur lissé peut s'écrire alors :

$$\dot{F}_n(x) = \frac{1}{a_n} \int K\left(\frac{x-t}{a_n}\right) \widehat{F}_n(t)\,\mathrm{d}t. \tag{2}$$

Dans cette Note, nous établissons un résultat du type Chung–Smirnov pour l'estimateur lissé défini en (2).

1. Introduction

Let $(X_n)_{n\geqslant 1}$ be a sequence of iid rv with common df F. These rv are regarded as the lifetimes of the items under study. In many situations, due to possible withdrawals of the items from the life testing experimentation, the lifetimes may not be directly observable. Instead, we observe only censored lifetimes. That is, assuming that $(C_n)_{n\geqslant 1}$ is another sequence of iid censoring random variables with the df G. The observable rv are the n pairs $\{(Z_i, \delta_i),\ i = 1, 2, \ldots, n\}$, where $Z_i = \min\{X_i, C_i\}$ and $\delta_i = \mathbb{1}_{\{X_i \leqslant C_i\}}$.

For this model, the coherent estimator $\widehat{F}_n$ of the df F based on $(Z_i, \delta_i),\ i = 1, 2, \ldots, n$, is the Kaplan–Meier [5] estimator (KME) defined by

$$1 - \widehat{F}_n(x) = \begin{cases} \displaystyle\prod_{i:\ Z_{(i)} \leqslant x} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}} & \text{if } x < Z_{(n)}, \\ 0 & \text{if } x \geqslant Z_{(n)}, \end{cases} \tag{3}$$

where $Z_{(1)} \leqslant Z_{(2)} \leqslant \cdots \leqslant Z_{(n)}$ are the order statistics of Z_i and $\delta_{(1)}, \delta_{(2)}, \ldots, \delta_{(n)}$ are the corresponding δ_i. Clearly, the Z_i are iid with common df $H(x) = 1 - (1 - F(x))(1 - G(x))$, and the uncensored model is the special case of the censored model with $G = 0$. Since the litterature about KME is huge, the interested reader can consult Gill [4]. Most notably, Breslow and Crowley [1] proved that the sequence of product-limit processes $\sqrt{n}(\widehat{F}_n(x) - F(x))$ converges weakly to a Gaussian process with zero mean and variance function

$$d(x) = \int_{-\infty}^{x} \frac{\mathrm{d}F(u)}{(1 - F(u))^2(1 - G(u))}. \tag{4}$$

If we suppose that the df F is absolutely continuous, it is natural to consider a smooth estimator $\dot{F}_n$ of F rather than the classical KME $\widehat{F}_n$ which is a step function. Such estimators arise quite naturally as integrals of nonparametric density estimators defined by (1). The smoothed KME can be represented by $\dot{F}_n(x) = \int_{-\infty}^{x} \hat{f}_n(t)\,\mathrm{d}t$, which can be rewritten, using integration by parts and Fubini's theorem, as follows

$$\dot{F}_n(x) = \frac{1}{a_n} \int K\left(\frac{x-t}{a_n}\right) \widehat{F}_n(t)\,\mathrm{d}t. \tag{5}$$

If no censoring is present, the Chung–Smirnov property has been obtained by Winter [9] for the upper bound and Degenhardt [3] for the lower bound. To the best of our knowledge, the problem of the smooth estimation in the censored case has been studied only for the Nelson–Aalen estimator using martingale theory and the recent result of Lemdani and Ould-Saïd [6] in which they prove that the asymptotic performance of $\dot{F}_n$ is better than the classical

KME $\widehat{F}_n$ in the sense of relative deficiency using the MSE criterion. There are few results about a smoothed df under censorship model. This is the goal of this Note.

For any df W, let $T_W = \sup\{x;\ W(x) < 1\}$. If $T_G < T_F$ the data over T_G is possible but cannot be observed. In this case we cannot estimate $F(x)$ for $x > T_G$.

2. Result

We recall that the Chung–Smirnov property (CSP) gives the rate of the convergence. We say that F_n satisfies the CSP if

$$\limsup_{n\to+\infty}\left(\frac{2n}{\log\log n}\right)^{1/2}\|F_n - F\| = C, \tag{6}$$

where $C = 1$ in the uncensored case (smoothed or not) and $\|f\| = \sup_x |f(x)|$.

We will establish the same result as (6) for smoothed Kaplan–Meier estimator $\dot{F}_n$.

The following conditions over the bandwidth $\{a_n\}$ and the kernel K will be needed.

$$a_n \downarrow 0, \quad na_n \uparrow \infty, \quad \frac{\log(1/a_n)}{na_n} \to 0 \ (a_n \text{ is not too small}), \quad \frac{\log(1/a_n)}{\log\log n} \to +\infty \ (a_n \text{ is not too big}), \tag{7}$$

$$K \text{ is with compact support } [-1,1], \quad \int uK(u)\,\mathrm{d}u = 0 \quad \text{and} \quad \int u^2K(u)\,\mathrm{d}u < +\infty. \tag{8}$$

In our approach, a strong representation of KME due to Major and Rejtö [7] plays an important role. Let $H^{un}(t) = \mathbb{P}[Z_1 \leqslant t,\ \delta_1 = 1]$ and $H^c(t) = \mathbb{P}[Z_1 \leqslant t, \delta_1 = 0]$, Major and Rejtö [7] established that, for $x < T_H$,

$$\widehat{F}_n(x) - F(x) = \big(1 - F(x)\big) \times \frac{1}{n}\sum_{i=1}^{n}\psi_i(x) + r_n(x) \tag{9}$$

and

$$\widehat{F}_n(x) - F(x) = \big(1 - F(x)\big) \times \frac{1}{\sqrt{n}}\Psi_n(x) + R_n(x), \tag{10}$$

where

$$\psi_i(x) = \frac{\mathbb{1}_{\{Z_i\leqslant x,\delta_i=1\}} - H^{un}(x)}{1 - H(x)} + \int_{-\infty}^{x}\frac{\mathbb{1}_{\{Z_i\leqslant y\}} - H(y)}{(1-H(x))^2}\,\mathrm{d}H^{un}(y) - \int_{-\infty}^{x}\frac{\mathbb{1}_{\{Z_i\leqslant y,\delta_i=1\}} - H^{un}(y)}{(1-H(x))^2}\,\mathrm{d}H(y) \tag{11}$$

and $\Psi_n(x)$ is a Gaussian process defined by

$$\Psi_n(x) = \frac{B_n(H^{un}(x))}{1 - H(x)} + \int_{-\infty}^{x}\frac{B_n(H^{un}(y)) - B_n(1 - H^c(y))}{(1-H(x))^2}\,\mathrm{d}H^{un}(y) - \int_{-\infty}^{x}\frac{B_n(H^{un}(y))}{(1-H(x))^2}\,\mathrm{d}H(y) \tag{12}$$

with $B_n(s),\ 0 \leqslant s \leqslant 1$, being a Brownian bridge. Moreover, the remainder terms in (9) and (10) satisfy, with probability one,

$$\mathbb{P}\left(\sup_{x\leqslant T} n\big|r_n(x)\big| > x + \frac{C}{\delta}\right) \leqslant \kappa\,\mathrm{e}^{-\lambda x\delta^2} \tag{13}$$

for all $x > 0$, where $C,\ \kappa > 0$ and $\lambda > 0$ are universal constants, and

$$\sup_{x\leqslant T}\big|R_n(x)\big| = \mathrm{O}\left(\frac{\log^2 n}{n}\right), \quad T < T_H. \tag{14}$$

Let $\|f\|_b = \sup_{x\leqslant b}|f(x)|$, our main result is the following:

Theorem 2.1. *Under the conditions* (7) *and* (8), *if F has density f having bounded first derivative f′ and if a_n is such that*

$$0 < a_n < 1 \quad \text{and} \quad \left(\frac{n}{\log\log n}\right)^{1/2} a_n^2 \longrightarrow 0 \quad \text{as } n \to +\infty, \tag{15}$$

then for all $T < T_H$

$$\limsup_{n\to+\infty}\left(\frac{2n}{\log\log n}\right)^{1/2} \|\dot{F}_n - F\|_T = C_{F,G} \quad a.s., \tag{16}$$

where $C_{F,G} = \sup_{t\leqslant T}(1-F(x))(d(x))^{1/2}$, $d(x)$ is defined in (4).

Remarks.

1. We pointed out that we have the same constant in (16) as in Csörgö and Horvath [2] for the maximal deviation between the KME's and the df F.
2. Our assumptions are the same as in Degenhardt [3] and our result extends his lower bound.
3. If no censoring is present, the problem is solved using some results on the empirical processes, which are not available in the censoring case. In our case, we use the strong approximation of the KME due to Major and Rejtö [7] to draw up the result.

Proof. First we prove the upper bound. Let $c_n = (2n/(\log\log n)^{1/2})$. By a Taylor expansion of order 2 and using (9), we get

$$c_n\big(\mathbb{E}\big[\dot{F}_n(x)\big] - F(x)\big) = \frac{c_n a_n^2}{2}\int u^2 f'(\xi_{u,x})K(u)\,\mathrm{d}u + c_n\mathbb{E}\left(\int r_n(x-a_nu)K(u)\,\mathrm{d}u\right), \tag{17}$$

where $\xi_{u,x}$ lies between x and $x - a_nu$. Now we have the following inequality

$$\left|c_n a_n^2 \int u^2 f'(\xi_{u,x})K(u)\,\mathrm{d}u\right| \leqslant \|f'\| c_n a_n^2 \int u^2 K(u)\,\mathrm{d}u. \tag{18}$$

Using the fact that, for all positive rv X, $\mathbb{E}(X) = \int_0^{+\infty}[1-F(u)]\,\mathrm{d}u$ and (13), we can prove that $\mathbb{E}[\sup_{u\leqslant T}|r_n(u)|] = \mathrm{O}(\frac{1}{n})$. Then the second term of the rhs of (17) is bounded by c_n/n which tends to zero.

Since $c_n a_n^2 \to 0$, from (8) it follows that

$$c_n\big\{\mathbb{E}\big[\dot{F}_n(x)\big] - F(x)\big\} \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{19}$$

Now, for the variance term, we prove that

$$\limsup_{n\to+\infty} c_n\|\dot{F}_n - \mathbb{E}\dot{F}_n\|_T \leqslant \limsup_{n\to+\infty}\left[c_n\|\widehat{F}_n - F\|_T + \mathrm{O}\left(\frac{c_n}{n}\right)\right].$$

By Csörgö and Horvath ([2], p. 413), one has

$$\limsup_{n\to+\infty} c_n\|\dot{F}_n - \mathbb{E}\dot{F}_n\|_T \leqslant C_{F,G}. \tag{20}$$

The result of the upper bound now follows by the triangle inequality and (19).

For the lower bound, using the reversed triangle inequality we get

$$\limsup_{n\to+\infty} c_n\|\dot{F}_n - F\|_T \geqslant \limsup_{n\to+\infty}\sup_{x\leqslant T}\frac{c_n}{\sqrt{n}}\left|\int_{-1}^{1}\Gamma_2 K(u)\,\mathrm{d}u\right| - \limsup_{n\to+\infty}\sup_{x\leqslant T}\frac{c_n}{\sqrt{n}}\left|\int_{-1}^{1}(\Gamma_1+\Gamma_3)K(u)\,\mathrm{d}u\right|,$$

where $\Gamma_1 = \beta_n(F(x-a_nu)) - \beta_n(F(x))$, $\Gamma_2 = \beta_n(F(x))$, $\Gamma_3 = F(x-a_nu) - F(x)$ and $\beta_n(F(x)) = \sqrt{n}(\widehat{F}_n(x) - F(x))$. Applying again Csörgö and Horvath [2], we have

$$\limsup_{n\to+\infty} \sup_{x\leqslant T} \frac{c_n}{\sqrt{n}} \left| \int_{-1}^{1} \Gamma_2 K(u)\,\mathrm{d}u \right| = \limsup_{n\to+\infty} c_n \|\widehat{F}_n - F\|_T = C_{F,G}. \tag{21}$$

By a Taylor expansion of F and (15), it is easy to see that the term in Γ_3 tends to zero.

Now, for Γ_1 it can be split as follows

$$\begin{aligned}
\Gamma_1 = {} & \big(1-F(x-a_nu)\big)\left\{ \frac{B_n(H^{un}(x-a_nu))}{1-H(x-a_nu)} + \int_{-\infty}^{x-a_nu} \frac{B_n(H^{un}(y)) - B_n(1-H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H^{un}(y) \right. \\
& \left. - \int_{-\infty}^{x-a_nu} \frac{B_n(H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H(y) \right\} - \big(1-F(x)\big)\left\{ \frac{B_n(H^{un}(x))}{1-H(x)} \right. \\
& \left. + \int_{-\infty}^{x} \frac{B_n(H^{un}(y)) - B_n(1-H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H^{un}(y) - \int_{-\infty}^{x} \frac{B_n(H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H(y) \right\} \\
& + \sqrt{n}\big(R_n(x-a_nu) - R_n(x)\big).
\end{aligned}$$

A Taylor expansion of $1-F(x-a_nu)$ in neighbourhood of x yields

$$\begin{aligned}
\Gamma_1 = {} & \big(1-F(x)\big)\left\{ \frac{B_n(H^{un}(x-a_nu))}{1-H(x-a_nu)} - \frac{B_n(H^{un}(x))}{1-H(x)} \right\} \\
& + \big(1-F(x)\big)\left\{ \int_{x}^{x-a_nu} \frac{B_n(H^{un}(y)) - B_n(1-H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H^{un}(y) + \int_{x-a_nu}^{x} \frac{B_n(H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H(y) \right\} \\
& + \big\{-a_nuf(\xi_{u,x}) + \mathrm{o}(a_n)\big\}\left\{ \frac{B_n(H^{un}(x-a_nu))}{1-H(x-a_nu)} + \int_{-\infty}^{x-a_nu} \frac{B_n(H^{un}(y)) - B_n(1-H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H^{un}(y) \right. \\
& \left. - \int_{-\infty}^{x-a_nu} \frac{B_n(H^{un}(y))}{(1-H(y))^2}\,\mathrm{d}H(y) \right\} + \sqrt{n}\big(R_n(x-a_nu) - R_n(x)\big) := \Gamma_{11} + \Gamma_{12} + \Gamma_{13} + \Gamma_{14}.
\end{aligned} \tag{22}$$

By the fact that B_n satisfies

$$\limsup_{n\to+\infty} \frac{\|B_n\|}{\sqrt{\log\log n}} < \infty \quad \text{a.s.} \tag{23}$$

and (8) we have $\limsup_{n\to+\infty} \sup_{x\leqslant T} \frac{c_n}{\sqrt{n}} \int_{-1}^{1} |\Gamma_{12} + \Gamma_{13}| K(u)\,\mathrm{d}u = 0$. Now for Γ_{11}, we have the following inequality

$$\begin{aligned}
& \limsup_{n\to+\infty} \sup_{x\leqslant T} \frac{c_n}{\sqrt{n}} \int_{-1}^{1} \big(1-F(x)\big) \left| \frac{B_n(H^{un}(x-a_nu))}{1-H(x-a_nu)} - \frac{B_n(H^{un}(x))}{1-H(x)} \right| K(u)\,\mathrm{d}u \\
& \leqslant \limsup_{n\to+\infty} \sup_{x\leqslant T} \frac{c_n}{\sqrt{n}} \int_{-1}^{1} \left\{ \frac{|B_n(H^{un}(x-a_nu)) - B_n(H^{un}(x))|}{1-H(x-a_nu)} \right\} K(u)\,\mathrm{d}u
\end{aligned}$$

$$+\limsup_{n\to+\infty}\sup_{x\leqslant T}\frac{c_n}{\sqrt{n}}\left|B_n\left(H^{un}(x)\right)\right|\int_{-1}^{1}\left|\frac{1}{1-H(x-a_nu)}-\frac{1}{1-H(x)}\right|K(u)\,\mathrm{d}u:=\Gamma'_{11}+\Gamma''_{11}. \tag{24}$$

Making use of (23), we can check that the term Γ''_{11} tends to zero. All what is left to be shown, is that Γ'_{11} is asymptotically negligible.

Let $\omega_n(a_n)=\sup_{|t-s|\leqslant a_n}|B_n(t)-B_n(s)|$ the continuity modulus of B_n.

By (7), the result of Stute [8] holds for B_n. That is, we have

$$\omega_n(a_n)=\mathrm{O}\left(\left(a_n\ln\frac{1}{a_n}\right)^{1/2}\right)\quad\text{a.s.} \tag{25}$$

Now by a Taylor expansion of $H^{un}(x-a_nu)$, the integral term of Γ'_{11} can be bounded by $\omega_n(a_n\|f\|)M$, where $M=\sup_{y\leqslant T+\varepsilon}\frac{1}{1-H(y)}$ for some $\varepsilon>0$ such that $T+\varepsilon<T_H$.

Now, using the following dominating sequence $\tilde{a}_n=(\frac{\log\log n}{n})^{1/4}$ which satisfy (15), we can check that, for all sequence a_n

$$\limsup_{n\to+\infty}\omega_n(a_n)=0\quad\text{a.s.} \tag{26}$$

Finally, the term Γ_{14} in (22) is of order $\frac{\log^2 n}{n}$ uniformly on $]-\infty,T]$, we have

$$\limsup_{n\to+\infty}\sup_{x\leqslant T}c_n\int_{-1}^{1}\left|R_n(x-a_nu)-R_n(x)\right|K(u)\,\mathrm{d}u=0. \tag{27}$$

Now (19)–(22), (26) and (27) allow us to conclude for the lower bound and the result.

Acknowledgements

The authors would like to thank the two anonymous referees for criticisms and remarks, in particular one of the referees for his very careful advice, which improved the presentation of this Note.

References

[1] N. Breslow, J. Crowley, A large sample study of the life table and product-limit estimates under random censorship, Ann. Statist. 2 (1974) 437–443.
[2] S. Csörgö, L. Horvath, The rate of strong uniform consistency for the product-limit estimator, Z. Warsch. Verw. Gebiete 62 (1983) 411–426.
[3] H.J.A. Degenhardt, Chung–Smirnov property for perturbed distribution function estimators, Statist. Probab. Lett. 16 (1993) 97–101.
[4] R.D. Gill, Lectures on Survival Analysis, in: Lectures on Probability Theory, Vol. 1581, Springer-Verlag, 1994.
[5] E.M. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc. 53 (1958) 457–481.
[6] M. Lemdani, E. Ould-Saïd, Relative deficiency of the Kaplan–Meier estimator with respect to smoothed estimator, Math. Methods Statist. 10 (2001) 215–234.
[7] P. Major, L. Rejtö, Strong embedding of the estimator of the distribution function under random censorship, Ann. Statist. 16 (1988) 1113–1132.
[8] W. Stute, The oscillation behaviour of empirical processes, Ann. Probab. 10 (1982) 86–107.
[9] B.B. Winter, Convergence rate of perturbed empirical distribution functions, J. Appl. Probab. 16 (1979) 163–173.