☑成果報告

行政院國家科學委員會補助專題研究計畫□期中進度報告

雙設限資料下存活函數之無母數估計值

計畫類別：☑個別型計畫　□整合型計畫
計畫編號：NSC　99－2118－M－029－003－MY2
執行期間：98 年 08 月 01 日至 100 年 07 月 31 日

執行機構及系所：東海大學統計系

計畫主持人：沈葆聖
共同主持人：
計畫參與人員：

成果報告類型(依經費核定清單規定繳交)：☐精簡報告　☑完整報告

本計畫除繳交成果報告外，另須繳交以下出國心得報告：
☐赴國外出差或研習心得報告
☐赴大陸地區出差或研習心得報告
☐出席國際學術會議心得報告
☐國際合作研究計畫國外研究報告

處理方式：除列管計畫及下列情形者外，得立即公開查詢
☐涉及專利或其他智慧財產權，☐一年□二年後可公開查詢

中　華　民　國 100 年 08 月 03 日

Nonparametric Estimators of the Survival Function with Twice Censored Data

Pao-sheng Shen

Department of Statistics
Tunghai University, Taichung, Taiwan, 40704
psshen@thu.edu.tw

Abstract

Patilea and Rolin (2006, *The Annals of Statistics*, **34(2)**, 925-938.) proposed a product-limit estimator of the survival function for twice censored data. In this article, based on a modified self-consistent (MSC) approach, we propose an alternative estimator, the MSC estimator. The asymptotic properties of the MSC estimator are derived. A simulation study is conducted to compare the performance between the two estimators. Simulation results indicate the MSC estimator outperforms the product-limit estimator and its advantage over the product-limit estimator can be very significant when right censoring is heavy.

Key Words: left-censored; right-censored; modified self-consistent.

1. Introduction

Consider the example of a reliability system which consists of three components E_1, E_2 and E_3, with E_1 and E_2 in series and E_3 is parallel with this series system. Let T, U and L denote the life times of the three components E_1, E_2 and E_3, respectively. Suppose that T, U and L are independent to one another and when the system fails we are able to determine which component failed at the same time as the system. When the system fails, we can only observe a lifetime variable $X = \max[\min(T, U), L]$ and an indicator variable δ with $\delta = 1$ if $L < T \leq U$, $\delta = 2$ if $L < U < T$ and $\delta = 3$ if $\min(T, U) \leq L$. We say that X is a twice censored observation of T. Note that twice censoring at first glance is the same as double censoring, where we also only observe a lifetime variable $X = \max[\min(T, U), L]$. However, the two schemes turns out to be quite distinct ideas. As an example of double censorship, Turnbull (1974) refers to a study of African infant precocity by Leiderman et al. (1973). A sample of 65 children were considered and each child was tested monthly to see if he (or she) had learned to accomplish certain tasks. The time from birth to the learning time was the variable of interest (denoted by T). In their analysis, double censoring occurred due to late

entries (i.e. left censoring: the child had already learned the skills before entering the study) and loss to follow-up (i.e. right censoring: the child had dropped out or not acquired the skill by the end of study). Hence, under double censoring scheme, L and U are dependent on each other with $P(L < U) = 1$, and we only observe a lifetime variable $X = \max[\min(T, U), L]$ and an indicator variable δ with $\delta = 1$ if $L < T \leq U$, $\delta = 2$ if $U < T$ and $\delta = 3$ if $T \leq L$.

Let $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ denote the observed sample and $Y_1 < Y_2 < \cdots < Y_J$ be the distinct values in increasing order of X_i's. For component E_1, let $d_j = \sum_{i=1}^{n} I_{[X_i = Y_j; \delta_i = 1]}$, denote the number of failure times at Y_j. Similarly, define $c_j = \sum_{i=1}^{n} I_{[X_i = Y_j; \delta_i = 2]}$ and $e_j = \sum_{i=1}^{n} I_{[X_i = Y_j; \delta_i = 3]}$ for components E_2 and E_2, respectively. Let F, Q, and G denote the distribution functions of T, U and L, respectively. Under this model, Patilea and Rolin (2006) proposed a product-limit estimator $\hat{S}_P(t)$ of $S_F(t) = 1 - F(t) = P(T > t)$ as follows:

$$\hat{S}_P(t) = \prod_{Y_i \leq t} \left\{ 1 - \frac{d_i}{n\hat{G}_n(Y_i) - R_n(Y_{i-1})} \right\},$$

where $R_n(x) = \sum_{j=1}^{J} I_{[Y_j \leq x]}$ and

$$\hat{G}_n(x) = \prod_{Y_j \geq x}^{J} \left(1 - \frac{e_j}{R_n(Y_j)} \right).$$

Note that $\hat{G}_n(x)$ is the product-limit estimator of $G(x-) = P(L < x)$ and $n^{-1}R_n(x-)$ is the empirical function of $P(X < x) = P(L < x)P(\min(T, U) < x)$. Hence, $\hat{G}_n(Y_i) - n^{-1}R_n(Y_{i-1})$ is a consistent estimator of the probability $P(L < t)P(\min(U, T) \geq t)$.

Let $a_F = \sup\{t : F(t) = 0\}$ and $b_F = \inf\{t : F(t) = 1\}$ denote the left and right support of T, respectively. Similarly, define a_Q and b_Q for U and a_G and b_G for L. Let $(D, ||\cdot||)$ be the Banach space of all real valued functions defined on $(0, \infty)$ which are right-continuous and have left limits at $t \leq \infty$, and $(D[a, b], ||\cdot||)$ the restrictions of $h \in D$ on $[a, b] \subset (0, \infty)$. Under the condition $F(a_F) = Q(a_Q) = 0$, Patilea and Rolin (2006) established the strong convergence of $\hat{S}_P$, i.e. $\sup_{a_F \leq t \leq b_F} |\hat{S}_P(t) - S_F(t)| \to 0$ almost surely. Let $\tilde{G}(t) = P(X \leq t, \delta = 3)$, $\tilde{Q}(t) = P(X \leq t, \delta = 2)$, $\tilde{F}(t) = P(X \leq t, \delta = 1)$, and $\tilde{H}(t) = \tilde{F}(t) + \tilde{Q}(t) + \tilde{G}(t)$. Let $\tau > \tilde{a}_F$ such that $\tilde{F}(\tau-) + \tilde{Q}(\tau-) < 1$. Under the condition $\int_{(a_{\tilde{F}}, \infty]} \tilde{G}(du)/[\tilde{H}(u)]^2 < \infty$, they also established the weak convergence of $\hat{S}_P(t)$ in $D[a_{\tilde{F}}, \tau]$. Note that when F, Q and G are continuous, $a_G \leq \min(a_F, a_Q)$ and $b_G = b_F = b_Q$, we have $a_{\tilde{F}} = a_F$ and $\tau = b_F$. In this case, the strong consistency and weak convergence of $\hat{S}_P(t)$ hold on the set $[a_F, b_F]$.

In Section 2, based on a modified self-consistent (MSC) approach, we propose an alternative estimator of $S_F(t)$, the MSC estimator (denoted by $\hat{S}_n$). The asymptotic properties of the MSC estimator are derived. In Section 3, a simulation is conducted to compare the performance between the two estimators, $\hat{S}_P$ and $\hat{S}_n$.

2. The Proposed Estimator

First, we shall demonstrate that the product-limit estimator $\hat{S}_P$ can be expressed as inverse-probability-weighted (IPW) average. For random censoring model, Satten and Datta (2001) showed that the Kaplan-Meier (1958) estimator of $F(t)$ can be expressed as an IPW average (see Robins (1993, 2000)). For the univariate random truncation and censoring model, Shen (2003) showed that the truncation nonparametric maximum likelihood estimator (NPMLE) (see Woodroofe (1985)) and the censoring-truncation NPMLE (see Wang (1987)) of survival function can also be expressed as IPW averages. For the twice censored data described in Section 1, the following arguments provide the motivation for using an IPW estimator.

Consider the subdistribution function

$$\tilde{F}(t) = P(L < T \leq U, T \leq t) = \int_{a_F}^{t} G(u-)S_Q(u-)F(dx),$$

where $S_Q(u) = 1 - Q(u)$. Thus, we have $F(dx) = [G(x-)S_Q(x-)]^{-1}\tilde{F}(dx)$. When G and S_Q are known, $S(t)$ can be estimated by

$$n^{-1}\sum_{i=1}^{n} \frac{I_{[X_i>t,\delta_i=1]}}{G(X_i-)S_Q(X_i-)},$$

where $I_{[.]}$ is the indicator function. Let $H(x) = P(\min(T, U) \leq x)$. Similar to $\hat{G}_n(x)$, we can estimate $H(x-)$ using the following product-limit estimator:

$$\hat{H}_n(x) = \prod_{Y_j \geq x} \left(1 - \frac{d_j + c_j}{R_n(Y_j)}\right).$$

Note that $1 - \hat{H}_n(x)$ is a consistent estimator of $S_F(x-)S_Q(x-)$. Hence, given $S_F(x)$, the survival function $S_Q(x-)$ can be estimated by $[1 - \hat{H}_n(x)]/S_F(x-)$. Therefore, an IPW estimator of $S_F(t)$ can be obtained by simultaneous solving the following two equations:

$$\hat{S}_W(t) = n^{-1}\sum_{j=1}^{J} \frac{d_j I_{[Y_j>t]}}{\hat{G}_n(Y_j)\hat{S}_{QW}(Y_j-)} \tag{1}$$

and

$$\hat{S}_{QW}(t-) = [1 - \hat{H}_n(t)]/\hat{S}_W(t-). \tag{2}$$

When F, Q and G are continuous, it is easy to show that $\hat{G}_n(Y_i)\hat{H}_n(Y_i) = n^{-1}R_n(Y_{i-1})$. Hence, $\hat{G}_n(Y_i) - n^{-1}R_n(Y_{i-1}) = \hat{G}_n(Y_i)[1 - \hat{H}_n(Y_i)]$. The jump in $\hat{S}_P$ at time Y_i is given by $\hat{S}_P(Y_{i-1})d_i/\{n\hat{G}_n(Y_i)[1 - \hat{H}_n(Y_i)]\}$. If we replace $\hat{S}_{QW}(Y_j-)$ in (1) by $\hat{S}_{Qp}(Y_j-) = [1 - \hat{H}_n(Y_j)]/\hat{S}_P(Y_{j-1})$, it follows that the jump in $\hat{S}_W$ at time Y_i is equal to that of $\hat{S}_P$. Hence, the product-limit estimator $\hat{S}_P$ can also be obtained by simultaneously solving the equations (1) and (2). By (1) and (2), it follows that the product-limit estimator $\hat{S}_P$ satisfies the following modified self-consistent equation:

$$\hat{S}_P(t) = \int_t^\infty \frac{\tilde{F}_n(dx)}{\frac{\hat{G}_n(x)-n^{-1}R_n(x-)}{\hat{S}_P(x-)}}.$$

Next, using the approach similar to the case of doubly censored data (see Turnbull (1974), Tsai and Crowley (1985), Chang and Yang (1987)), we propose a modified self-consistent (MSC) estimator of S_F. Consider the subdistribution function $\tilde{Q}(t) = P(L < U < T, U \leq t) = \int_{a_F}^t G(u-)S_F(u)Q(du)$. When G and S_F are known, $S_Q(t)$ can be estimated by

$$n^{-1}\sum_{i=1}^n \frac{I_{[X_i>t,\delta_i=2]}}{G(X_i-)S_F(X_i)}.$$

Let $W(t) = P(X > t)$ and $S_G(t) = 1 - G(t)$. Then

$$W(t) = S_G(t) + S_F(t)S_Q(t) - S_F(t)S_G(t)S_Q(t). \tag{3}$$

Note that equation (3) can be rewritten as

$$S_F(t) = W(t) - S_G(t) + S_F(t)Q(t) + S_F(t)S_G(t)S_Q(t).$$

Let $\tilde{W}_n$ and $\tilde{Q}_n$ denote the empirical survival and distribution functions of W and $\tilde{Q}$, respectively. Now, we require the estimators of S_F, Q and S_Q (denoted by $\hat{S}_n$, $\hat{Q}_n$ and $\hat{S}_Q$, respectively) to relate $\tilde{Q}_n$. Imposing the condition $\hat{S}_Q(0) = 1$, we have $\hat{Q}_n(t) = \int_0^t \frac{1}{\hat{G}_n(u)\hat{S}_n(u)}\tilde{Q}_n(du)$ and $\hat{S}_Q(t) = 1-\hat{Q}_n(t)$. Thus, a MSC estimator $\hat{S}_n(t)$ can be obtained by solving the following equation:

$$\hat{S}_n(t) = \tilde{W}_n(t) - \hat{S}_{G_n}(t) + \hat{S}_n(t)\int_0^t \frac{\tilde{Q}_n(du)}{\hat{G}_n(u)\hat{S}_n(u)} + \hat{S}_n(t)\hat{S}_{G_n}(t)\left[1 - \int_0^t \frac{\tilde{Q}_n(du)}{\hat{G}_n(u)\hat{S}_n(u)}\right], \tag{4}$$

where $\hat{S}_{G_n}(t) = 1 - \hat{G}_n(t)$.

Note that for doubly censored data (i.e. $P(L < U) = 1$), equation (4) is reduced to

$$\hat{S}_n(t) = \tilde{W}_n(t) + \hat{S}_n(t) \int_0^t \frac{\tilde{Q}_n(du)}{\hat{S}_n(u)} - [1 - \hat{S}_n(t)]\hat{S}_{G_n}(t). \tag{5}$$

In this case, since the subdistribution function $\tilde{G}(t) = P(X_i \leq t, \delta_i = 3) = \int_{a_G}^t [1 - S_F(x-)]G(dx)$. When S_F is known, $S_G(t)$ can be estimated by $\int_t^\infty [1/(1 - \hat{S}_n(t))]\tilde{G}_n(dt)$, where $\tilde{G}_n$ is the empirical function of $\tilde{G}(t)$. Replacing $\hat{S}_{G_n}$ in (5) with $\int_t^\infty [1/(1-\hat{S}_n(t))]d\tilde{G}_n(t)$, we obtain the following self-consistent estimation equation for doubly censored data (see (5.1) in Tsai and Crowley (1985) or (2.11) in Chang and Yang (1987)):

$$\hat{S}_n(t) = \tilde{W}_n(t) + \hat{S}_n(t) \int_0^t \frac{\tilde{Q}_n(du)}{\hat{S}_n(u)} - [1 - \hat{S}_n(t)] \int_t^\infty \frac{1}{1 - \hat{S}_n(u)} \tilde{G}_n(du). \tag{6}$$

Before we go into deriving the asymptotic properties of the MSC estimator $\hat{S}_n$, we briefly review the asymptotic properties of the self-consistent estimator $\hat{S}_n$ for doubly censored data (i.e. for the case $P(L < U) = 1$). Under assumptions (A) $P(L \leq t \leq U) > 0$ for $t \in (a_F, b_F)$, Gu and Zhang (1993) showed that $\sup_{t \in (a_F, b_F)} |\hat{S}_n(t) - S_F(t)| \to 0$ a.s.. Under assumptions (A) and (B) $\int_{\tau < S_F(u) < 1} \frac{Q(du)}{G(u) - Q(u)} + \int_{0 < S_F(u) < \tau} \frac{G(du)}{G(u) - Q(u)} < \infty$ for all $0 < \tau < 1$, Gu and Zhang (1993) obtained the asymptotic normality of $\hat{S}_n(t)$ on (a_F, b_F). Mykland and Ren (1996) (see Theorem 2) showed that the nonparametric maximum likelihood estimator (NPMLE) satisfies the equation (6) and provided an explicit sufficient and necessary condition for a self-consistent estimator to be the NPMLE. Theorem 2 of Mykland and Ren (1996) implies that the NPMLE is a self-consistent estimator. Their proof is based on the following likelihood function of (Y_j, d_j, c_j, e_j) $(j = 1, \ldots, J)$ for S_F:

$$L(S_F) = C \prod_{j=1}^J (S_F(Y_{j-1}) - S_F(Y_j))^{d_j} (S_F(Y_j))^{c_j} (1 - S_F(Y_j))^{e_j},$$

where C is the term that depends only on (L, U) and $S_F(Y_0) = 1$. However, for twice censored data, since $P(L < U) < 1$, the likelihood function for S_F is proportional to the function

$$L(S_F) = \prod_{j=1}^J (S_F(Y_{j-1}) - S_F(Y_j))^{d_j} (S_F(Y_j))^{c_j} [1 - S_F(Y_j) S_Q(Y_j)]^{e_j}.$$

Since the likelihood function involves both S_F and S_Q, it is not easy to derive the NPMLE of S_F. We briefly discuss the relationship between the NPMLE and self-consistent estimator.

Define $f_0(x) = P_{S_F}(X = x)$ and $f(x) = P_S(X = x)$. Given S_Q and G, we have

$$E_{S_F}[\log L(S)|X_i = x_i, \delta_i] = \sum_{i=1}^{n}\Bigg[I_{[\delta_i=1]}\log f(x_i) + I_{[\delta_i=2]}\frac{\sum_{x_j>x_i} f_0(x_j)\log f(x_j)}{S_F(x_i)G(x_i)}$$

$$+I_{[\delta_i=3]}\frac{S_Q(x_i)\sum_{x_j\le x_i} f_0(x_j)\log f(x_j)}{1-S_F(x_i)S_Q(x_i)} + I_{[\delta_i=3]}\frac{Q(x_i)\sum_{j=1}^{n} f_0(x_j)\log f(x_j)}{1-S_F(x_i)S_Q(x_i)}\Bigg].$$

Similar to the proof of Theorem 6 of Mykland and Ren (1996), given S_Q, we can show that the following modified self-consistent equation is asymptotically equivalent to the EM algorithm:

$$\hat{S}_E(t) = \tilde{W}_{1n}(t) + \tilde{W}_{2n}(t) + \hat{S}_E(t)\int_0^t \frac{\tilde{Q}_n(du)}{\hat{G}_n(u)\hat{S}_E(u)} + \hat{S}_E(t)\int_0^\infty \frac{Q(u)}{1-\hat{S}_E(u)S_Q(u)}\tilde{G}_n(du)$$

$$+\int_t^\infty \frac{S_Q(u)[\hat{S}_E(t)-\hat{S}_E(u)]}{1-\hat{S}_E(u)S_Q(u)}\tilde{G}_n(du),$$

where $\tilde{W}_{1n}(t)$ and $\tilde{W}_{2n}(t)$ are the empirical functions of $\tilde{W}_1(t) = P(X_i > t, \delta_i = 1)$ and $\tilde{W}_2(t) = P(X_i > t, \delta_i = 1)$, respectively. Note that the equation above is different from equation (4), which does not involve S_Q. For twice censored data, further research is required to establish the relationship between the NPMLE and SCE.

Next, we shall derive the asymptotic properties of the MSC estimator $\hat{S}_n$. The proof of the following Theorem is inspired from Gu and Zhang (1993), where they derived the asymptotic properties of the SCE of equation (6) for doubly censored data.

Theorem 1.

Suppose that G is continuous and

$$S_Q(t)G(t) > 0 \text{ holds on } (a_F, b_F). \tag{7}$$

Then, $\sup_{a_F<t<b_F} |\hat{S}_n(t) - S_F(t)| \to 0$ a.s..

Proof:

Similar to Theorem 1 of Gu and Zhang (1993), we shall first prove the uniqueness of the solution (4) (STEP A) and then prove the uniform consistency of $\hat{S}_n$ (STEP B).

STEP A: Uniqueness of the solution of (4)

The proof of uniqueness is similar to that of Lemma 1 of Gu and Zhang (1993) (see Appendix, page 619). First, since $\tilde{W}_n \to W$ and $\hat{G}_n \to G$ uniformly and $\hat{S}_n$ satisfies (4), $\hat{S}_{n_k}(t) \to S(t)$ for each t as $n_k \to \infty$ implies

$$S(t) = W(t) - S_G(t) + S(t) \int_0^t \frac{\tilde{Q}(du)}{G(u)S(u)} + S(t)S_G(t) \int_t^\infty \frac{\tilde{Q}(du)}{G(u)S(u)}. \tag{8}$$

Let S be a [0,1]-valued nonincreasing function with left support a_F and right support b_F. Let h be a function such that

$$h(t)K(t) = -\int_{u \le t} \frac{S(t)}{G(u)S(u)} h(u)Q(du) - \int_{u>t} \frac{S(t)S_G(t)}{G(u)S(u)} h(u)Q(du), \tag{9}$$

where $K(t) = S_Q(t)G(t)$. Suppose that (7) holds. We shall show that $h(t) = 0$ for all $t \in (a_F, b_F)$ by setting $h(t) = S(t) - S_F(t)$. Assume $h(t_0) > 0$ and $0 < S(t_0) < 1$ at some point t_0. Our goal is to establish a contradiction.

Define

$$g(t) = -\int_{u \le t} \frac{h(u)}{G(u)S(u)} Q(du) - \int_{u>t} \frac{h(u)S_G(t)}{G(u)S(u)} Q(du).$$

Note that both $K(t)$ and $g(t)$ are right continuous and their definitions are different from that of Gu and Zhang (1993), where they were defined as $K(t) = S_Q(t) - S_G(t)$ and

$$g(t) = -\int_{u \le t} \frac{h(u)}{S(u)} Q(du) + \int_{u>t} \frac{h(u)}{1 - S(u)} G(du).$$

By (9), we have $K(t)h(dt+) = g(t)S(dt+)$ and $K(t-)h(dt) = g(t-)S(dt)$, where $h(dt) = h(t) - h(t-)$ and $h(dt+) = h(t+) - h(t)$. Define $t_1 = \sup\{a_F < t \le t_0 : h(t) \le 0\}$, $t_2 = \inf\{t_0 \le t < b_F : h(t) \le 0\}$ and $\mathcal{H} = \{t : h(t) > 0, t_1 \le t \le t_2\}$. Then, $t_0 \in \mathcal{H}$, and $(t_1, t_2) \subset \mathcal{H} \subset [t_1, t_2]$. First,

$$g(dt) = -h(t)\left[\frac{Q(dt)}{G(t)S(t)} - \frac{S_G(t)Q(dt)}{G(t)S(t)}\right] \le 0 \quad on \quad \mathcal{H}. \tag{10}$$

To establish a contradiction, we need Step 1 as follows.

Step 1: Show that $g(t) = g(t-) = 0$ on $\mathcal{H}$.

By assumption (7) and the arguments of Lemma 1 of Gu and Zhang (1993) (see page 620), it follows that $g(t) = g(t-) = 0$ on $\mathcal{H}$.

Step 2: Find a contradiction.

Since $h(t) > 0$ on $\mathcal{H}$, by assumption (7), Step 1 and (10), we have $Q(dt) = 0$, $K(t) = K(t-) =$constant > 0 on $\mathcal{H}$. Hence, we have $h(dt) = h(dt+) = 0$ on $\mathcal{H}$, so that $h(t) = h(t_0) > 0$ on $\mathcal{H}$ and $t_0 \in \mathcal{H} = (t_1, t_2)$. Therefore, $h(t_1) \leq 0$ and $h(t_1+) = h(t_0) > 0$. Since $K(t_1) > 0$, it follows that $g(t_1+) < 0$, which is a contradiction to Step 1, i.e. $g(t_1+) = 0$.

By setting $h(t) = S(t) - S_F(t)$, it follows that $h(t) = 0$ for $t \in (a_F, b_F)$. The proof of the uniqueness is completed.

STEP B: Uniform consistency

By (4) all limit points of $\hat{S}_n$ must satisfy (8), by Helly-Bray selection theorem we have $\hat{S}_n(t) \to S_F(t)$ a.s. for $t \in (a_F, b_F)$. Let $\tilde{S}(t) = P(X > t, \delta = 1)$ and $\tilde{S}_n(t)$ be the empirical function of $\tilde{S}(t)$. If $S_F(dt) < 0$ then by (4),

$$\frac{\tilde{S}_n(dt)}{\hat{S}_n(dt)} \leq 1 - \int_{u<t-\epsilon} \frac{\tilde{Q}_n(du)}{\hat{G}_n(u)\hat{S}_n(u)} - [1 - \hat{G}_n(t)] \int_{u \geq t} \frac{\tilde{Q}_n(du)}{\hat{G}_n(u)\hat{S}_n(u)} \to \frac{\tilde{S}(dt)}{S_F(dt)}$$

as $n \to \infty$ and then $\epsilon \to 0+$, which implies $|\hat{S}_n(dt)| \geq (1-o(1))|S_F(dt)|$, since $\tilde{S}_n(dt)/\tilde{S}(dt) \to 1$. Hence, $\sup_{t \in (a_F, b_F)} |\hat{S}_n(t) - S_F(t)| \to 0$ a.s.

The proof is completed. □

In order to derive the asymptotic normality of $\sqrt{n}[\hat{S}_n(t) - S_F(t)]$, similar to Theorem 2 of Gu and Zhang (1993) (see page 613), we define four linear operators as follows. For any survival function S, let A_S, R_S, K and B_S be the linear operators defined by

$$(A_S h)(t) = -\int_{u \leq t} \frac{S(t)}{G(u)S(u)} h(u)Q(du) - \int_{u>t} \frac{S(t)[1-G(t)]}{G(u)S(u)} h(u)Q(du), \tag{11}$$

$$R_S = A_S - K, \ (Kh)(t) = K(t)h(t), \tag{12}$$

and

$$B_S(h^{(1)}, h^{(2)}, h^{(3)}, h^{(4)})(t) =$$

$$\sum_{j=1}^{3}[1-h^{(j)}(t)] - [1-h^{(4)}(t)] - \int_{u \leq t} \frac{S(t)}{S(u)h^{(4)}(u-)} h^{(2)}(du) - \int_{u>t} \frac{S(t)[1-h^{(4)}(t)]}{S(u)h^{(4)}(u-)} h^{(2)}(du). \tag{13}$$

By (11), we have

$$A_{\hat{S}_n}(\hat{S}_n - S_F) = B_{\hat{S}_n}(\tilde{F}, \tilde{Q}, \tilde{G}, G)(t) - \hat{S}_n(t) + K(t)[\hat{S}_n(t) - S_F(t)].$$

Since $\hat{S}_n(t) = B_{\hat{S}_n}(\tilde{F}_n, \tilde{Q}_n, \tilde{G}_n, \hat{G}_n)(t)$, $R_{\hat{S}_n}\xi_n = B_{\hat{S}_n}Z_n$, where $\xi_n = \sqrt{n}(\hat{S}_n - S_F)$ and

$$Z_n = (\sqrt{n}(\tilde{F}_n - \tilde{F}), \sqrt{n}(\tilde{Q}_n - \tilde{Q}), \sqrt{n}(\tilde{G}_n - \tilde{G}), \sqrt{n}(\hat{G}_n - G)).$$

Let $(D(a_F, b_F), ||\cdot||_F)$ be the Banach space of all real-valued functions defined on (a_F, b_F) which are right-continuous and have left-limit at $t < b_F$, where $||h(t)||_F = \sup_{a_F<t<b_F} h(t)$. Define Banach spaces $(D_K(a_F, b_F), ||\cdot||_K) = \{h : Kh \in D(a_F, b_F)\}$, $||h||_K = ||Kh||_F$, $(D_Z, ||\cdot||_Z) = \{h \in D \otimes D \otimes D \otimes D : B_{S_F}(h) \in D(a_F, b_F)\}$, $||(h^{(1)}, h^{(2)}, h^{(3)}, h^{(4)})||_Z = \sum_{j=1}^4 ||h^{(j)}||_F$. By (11), we have $B_{S_F}(\tilde{F}_n - \tilde{F}), (\tilde{Q}_n - \tilde{Q}), (\tilde{G}_n - \tilde{G}), (\hat{G}_n - G)) \in D(a_F, b_F)$ and

$$Z_n \xrightarrow{d} Z = (Z_1, Z_2, Z_3, Z_4) \text{ in } D_Z, \tag{14}$$

where $E[Z_i(t)] = 0$ $(i = 1, 2, 3, 4)$, $E[Z_1(t)Z_1(s)] = \tilde{F}(\max(t, s)) - \tilde{F}(t)\tilde{F}(s), \cdots$, $E[Z_4(t)Z_4(s)] = G(\max(t, s)) - G(t)G(s)$; and $E[Z_1(t)Z_2(s)] = -\tilde{F}(t)\tilde{Q}(s), \cdots$, $E[Z_3(t)Z_4(s)] = -\tilde{G}(t)G(s)$.

Next, we derive the asymptotic normality of $\sqrt{n}(\hat{S}_n - S_F)$. The proof of the following theorem is similar to that of Theorem 2 of Gu and Zhang (1993) (see page 617). The main idea of the proof is via strong continuity of linear operators indexed by survival functions in the metric space $\mathcal{F}_S = \{S : S - S_F \in D(a_F, b_F)\}$ with the distance $||S - S_F||_F$.

Theorem 2. Under the assumptions of Theorem 1 and

$$\int_{(a_F, b_F)} \frac{1}{G(t)} Q(dt) > 0, \tag{15}$$

Then, $R_{S_F}^{-1}$, the inverse of R_{S_F} in (14), exists as a bounded operator from $D(a_F, b_F)$ to $D_K(a_F, b_F)$, and

$$\sqrt{n}(\hat{S}_n(t) - S_F(t)) = \xi_n \xrightarrow{d} \xi = R_{S_F}^{-1} B_{S_F} Z \text{ in } D_K(a_F, b_F),$$

where Z is the Gaussian process in (14).

Proof:

Let $S_{T,m}$ $m \geq 1$ be a finite discrete survival function such that $||S_{T,m} - S||_F \to 0$. Let $Q_{U,m}$ be a finite discrete distribution function such that $||Q_{U,m} - Q||_F \to 0$. Note that the existence of $S_{T,m}$ and $Q_{U,m}$ is guaranteed by (7) and (15). Let h_m, g_m. $m \geq 1$, and g be functions in $D(a_F, b_F)$ such that $||g_m - g|| \to 0$ and $R_m h_m = g_m$, where $R_m = A_m - K_m$, and A_m, R_m, and K_m are defined as (11)-(13) with (S, Q, G) replaced by $(S_{T,m}, Q_{U,m}, \hat{G}_m)$.

Under assumption (15) we have

$$\lim_{\tau\to a_F+}\sup_m\left[\int_{1-\tau<S_F(u)<1} 1/\hat{G}_m(u)Q_{U,m}(du)+\int_{0<S_F(u)<\tau} 1/\hat{G}_m(u)Q_{U,m}(du)\right]=0. \qquad (16)$$

Note that condition (16) is different from condition (4.3) of Gu and Zhang (1993) (see Lemma 2, page 617), which is the following:

$$\lim_{\tau\to a_F+}\sup_m\left[\int_{1-\tau<S_F(u)<1} 1/[G_{U,m}(u)-Q_{U,m}]Q_{U,m}(du)+\int_{0<S_F(u)<\tau} 1/[G_{U,m}(u)-Q_{U,m}]G_{U,m}(du)\right]=0,$$

where $G_{U,m}$ be a finite discrete distribution function such that $||G_{U,m}-Q||_F\to 0$.

In our case, condition (16) is required for the existence of $h\in D_K(a_F,b_F)$ such that $||Kh_m-Kh||_F\to 0$ and $R_Sh=g$. Define

$$v_m^-=-\int_{u\le t}\frac{S_{T,m}(t)h_m(u)}{\hat{G}_m(u)S_{T,m}(u)}Q_{U,m}(du)$$

and

$$v_m^+=-\int_{u>t}\frac{S_{T,m}(t)S_G(t)h_m(u)}{\hat{G}_m(u)S_{T,m}(u)}Q_{U,m}(du).$$

Note that the definitions of v_m^- and v_m^+ are different from that of Gu and Zhang (1993), where they were defined as

$$v_m^-=-\int_{u\le t}\frac{S_{T,m}(t)h_m(u)}{\hat{S}_{T,m}(u)}Q_{U,m}(du)$$

and

$$v_m^+=-\int_{u>t}\frac{1-S_{T,m}(t)h_m(u)}{\hat{1}-S_{T,m}(u)}G_{U,m}(du).$$

By definition of v_m^+, v_m^- and $R_mh_m=g_m$, we have $K_mh_m=g_m+v_m^++v_m^-$.

Step 1: Existence of R_S^{-1} as a linear operator from $D(a_F,b_F)$ to $D_K(a_F,b_F)$ for all $S\in\mathcal{F}_S=\{S:S-S_F\in D(a_F,b_F)\}$.

A sequence of functions is totally bounded if every subsequence contains a uniformly convergent further subsequence. By (14), for a fixed $0<\tau_0<1$, both $v_m^+(t)I_{[S(t)\le\tau_0]}$ and $v_m^-(t)I_{[S(t)>\tau_0]}$ are totally bounded for the case $||K_mh_m||_F\le 1$. Further, since $S-S_F\in D(a_F,b_F)$, we have $||v_m^-(t)||_F=o_p(1)$ and $||v_m^+(t)||_F=o_p(1)$. Similar to the arguments of Steps 3 and 4 of Lemma 2 of Gu and Zhang (1993)), under condition (16), it follows that

there exists $h \in D_K(a_F, b_F)$ such that $||Kh_m - Kh||_F \to 0$, $R_S h = g$ and the solution h of $R_S h = g$ is unique. Define $h = R_S^{-1} g$. This completes the proof of Step 1.

Step 2: Strong continuity of $\{R_S^{-1}, S \in \mathcal{F}_S\}$.

Let $g_m^{'} \in D(a_F, b_F)$ and S_m in $\mathcal{F}_S$ be such that $||g_m^{'} - g||_F \to 0$ and $||S_m - S||_F \to 0$. Similar to Step 2 of Theorem 2 of Gu and Zhang (1993), it follows that $||K_m h_m - K R_S^{-1} g||_F \to 0$, which implies that $||K R_{S_{T,m}}^{-1} g_m^{'} - K R_S^{-1} g||_F \to 0$. Hence, we have the strong continuity. This completes the proof of Step 2.

Step 3: Strong continuity of $\{B_S, S \in \mathcal{F}_S\}$.

Let h be a simple function in D_Z. Since $S - S_F \in D(a_F, b_F)$, by (13), $B_S h \to B_{S_F} h$ in $D(a_F, b_F)$ as $||S - S_F||_F \to 0$. Since $||B_S||_F \leq 2$ and the collection of simple functions is dense in D_Z, we have the strong continuity. This completes the proof of Step 3.

Similar to the arguments of Steps 4 and 5 of Theorem 2 of Gu and Zhang (1993), it follows that $\sqrt{n}(\hat{S}_n(t) - S_F(t))$ converges in distribution in $D_K(a_F, b_F)$.

The proof is completed. □

3. Simulation Study

A simulation study is conducted to compare the performance between the two estimators $\hat{S}_P$ and $\hat{S}_n$. The T's are i.i.d. exponential distributed with scale parameter equal to 1, i.e. $F(x) = 1 - e^{-x}$ for $x > 0$. The U's are i.i.d. exponential distributed with scale parameters λ_q, i.e. $Q(x) = 1 - e^{-\lambda_q x}$ for $x > 0$. The L's are i.i.d. exponential distributed with scale parameters λ_g, i.e. $G(x) = 1 - e^{-\lambda_g x}$ for $x > 0$. The T, U and L are independent to one another. Then the variables X and δ are generated as described in Section 1. The goal is to estimate the survival function of T: $S_F(t) = p_f$, where p_f is chosen as $p_f = 0.75, 0, 5, 0.25$. The values of (λ_g, λ_q) are chosen as $(0.75, 1.0)$, $(30, 4.0)$, $(15, 1.0)$, $(2.0, 0.5)$, and $(4.0, 0.1)$. The sample sizes are chosen as 100 and 200. The replication is 1000 times. Tables 1 through 3 show the biases and root mean squared errors (denoted by rmse) of the two estimators for $S(0.29) = 0.75$, $S(0.69) = 0.5$ and $S(1.39) = 0.25$, respectively. Tables 1 through 3 also show the ratio of root mean squared errors of $\hat{S}_n$ to that of $\hat{S}_P$ (denoted by r. Further, Tables 1 through 3 also list the simulated proportions for $\delta = 1$, $\delta = 2$ and $\delta = 3$, denoted by p_1, p_2 and p_3, respectively.

Table 1. Simulation results for biases and rmse of $\hat{S}_P$ and $\hat{S}_n$, $S(0.29) = 0.75$

						$\hat{S}_P(0.29)$		$\hat{S}_n(0.29)$		
λ_g	λ_q	n	p_1	p_2	p_3	bias	rmse	bias	rmse	r
0.75	1.0	100	0.14	0.16	0.70	0.035	0.135	-0.032	0.112	0.83
0.75	1.0	200	0.14	0.16	0.70	0.034	0.091	-0.019	0.077	0.85
30	4.0	100	0.15	0.71	0.14	0.019	0.065	-0.003	0.060	0.92
30	4.0	200	0.15	0.71	0.14	-0.001	0.049	-0.010	0.044	0.90
15	1.0	100	0.42	0.44	0.14	0.010	0.053	-0.006	0.051	0.96
15	1.0	200	0.42	0.44	0.14	0.008	0.036	-0.009	0.037	1.03
2.0	0.5	100	0.43	0.16	0.41	0.022	0.082	-0.013	0.075	0.87
2.0	0.5	200	0.43	0.16	0.41	0.014	0.061	-0.010	0.055	0.90
4.0	0.1	100	0.72	0.10	0.18	0.024	0.065	-0.003	0.051	0.78
4.0	0.1	200	0.72	0.10	0.18	0.014	0.043	-0.003	0.037	0.86

Table 2. Simulation results for biases and rmse of $\hat{S}_P$ and $\hat{S}_n$, $S(0.69) = 0.50$

						$\hat{S}_P(0.69)$		$\hat{S}_n(0.69)$		
λ_g	λ_q	n	p_1	p_2	p_3	bias	rmse	bias	rmse	r
0.75	1.0	100	0.14	0.16	0.70	0.037	0.142	-0.017	0.128	0.90
0.75	1.0	200	0.14	0.16	0.70	0.019	0.095	-0.009	0.090	0.95
30	4.0	100	0.15	0.71	0.14	0.015	0.124	0.021	0.103	0.83
30	4.0	200	0.15	0.71	0.14	0.009	0.103	0.011	0.078	0.76
15	1.0	100	0.42	0.44	0.14	0.004	0.066	-0.005	0.067	1.02
15	1.0	200	0.42	0.44	0.14	0.005	0.045	-0.002	0.043	0.96
2.0	0.5	100	0.43	0.16	0.41	0.029	0.082	-0.003	0.076	0.93
2.0	0.5	200	0.43	0.16	0.41	0.016	0.054	-0.003	0.054	1.00
4.0	0.1	100	0.72	0.10	0.18	0.008	0.056	-0.005	0.052	0.93
4.0	0.1	200	0.72	0.10	0.18	0.011	0.043	-0.001	0.039	0.91

Table 3. Simulation results for biases and rmse of $\hat{S}_P$ and $\hat{S}_n$, $S(1.39) = 0.25$

						$\hat{S}_P(1.39)$		$\hat{S}_n(1.39)$		
λ_g	λ_q	n	p_1	p_2	p_3	bias	rmse	bias	rmse	r
0.75	1.0	100	0.14	0.16	0.70	0.028	0.105	0.007	0.097	0.92
0.75	1.0	200	0.14	0.16	0.70	0.011	0.075	-0.007	0.069	0.92
30	4.0	100	0.15	0.71	0.14	0.165	0.216	0.113	0.146	0.68
30	4.0	200	0.15	0.71	0.14	0.100	0.171	0.074	0.095	0.55
15	1.0	100	0.42	0.44	0.14	0.012	0.074	0.007	0.072	0.97
15	1.0	200	0.42	0.44	0.14	0.006	0.049	0.003	0.047	0.96
2.0	0.5	100	0.43	0.16	0.41	0.024	0.062	0.009	0.058	0.94
2.0	0.5	200	0.43	0.16	0.41	0.012	0.047	0.003	0.043	0.91
4.0	0.1	100	0.72	0.10	0.18	0.008	0.050	0.001	0.045	0.90
4.0	0.1	200	0.72	0.10	0.18	0.005	0.035	-0.003	0.035	1.00

Based on the results of Tables 1 through 3, we conclude that:

(i) For the estimation of $S(0.29) = 0.25$, the bias and standard deviation of $\hat{S}_n$ are smaller than that of $\hat{S}_n$ for most of the cases considered. In terms of rmse, the MSC estimator $\hat{S}_n$ outperforms the product estimator $\hat{S}_P$. The ratio of root mean squared errors of $\hat{S}_n$ to that of $\hat{S}_P$ ranges from 0.78 to 1.03.

(ii) For the estimation of $S(0.69) = 0.5$, the bias and standard deviation of $\hat{S}_n$ are smaller than that of $\hat{S}_P$ for most of the cases considered. In terms of rmse, the estimator $\hat{S}_n$ outperforms $\hat{S}_P$. The ratio of root mean squared errors of $\hat{S}_n$ to that of $\hat{S}_P$ ranges from 0.76 to 1.02.

(iii) For the estimation of $S(1.69) = 0.75$, the bias and standard deviation of $\hat{S}_n$ are smaller than that of $\hat{S}_P$ for all the cases considered. In terms of rmse, the estimator $\hat{S}_n$ outperforms $\hat{S}_P$. When right censoring is heavy (i.e. $p_2 = 0.71$), the bias and standard deviation of $\hat{S}_n$ are much smaller than that of $\hat{S}_P$. One explanation for the results is that $\tilde{F}_n$ on which $\hat{S}_P$ is based, is a function of the data with $X_i \leq t$ and $\delta_i = 1$ and $\tilde{Q}_n$, on which $\hat{S}_n$ is based, is a function of the data with $X_i \leq t$ and $\delta_i = 2$. The ratio of root mean squared errors of $\hat{S}_n$ to that of $\hat{S}_P$ ranges from 0.55 to 1.00.

4. Discussion

For twice censored data considered by Patilea and Rolin (2006), we have proposed an

alternative estimator, the MSC estimator, and established its asymptotic properties. Our simulation results indicate that the MSC estimator outpeforms the produt-limit estimator. The advantage of the MSC estimator over the product-limit estimator can be very significant when right censoring is heavy.

Reference

Chang, M. N. and Yang, G. L. (1987), Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *The Annals of Statistics* **15**, 1536-1547.

Gu, M. G. and Zhang, C. H. (1993), Asymptotic properties of self-consistent estimators based on doubly censored data. *The Annals of Statistics* **21**, 611-624.

Kaplan, E. L. and Meier, P. (1958), Nonparametric estimation from incomplete observations, *Journal of the American Statistical Assocication*, **53**, 457-481.

Leiderman, P. H., Babu, D., Kagia, J., Kraemer, H. C., and Leiderman, G. F. (1973). African infant precocity and some social influences during the first year. *Nature*, **242**, 247-249.

Mykland, P, A. and Ren, J. (1996), Algorithms for computing self-consistent and maximum likelihood estimators with doubly censored data. *The Annals of Statistics* **24**, 1740-1764.

Patilea, V. and Rolin, J.-M. (2006), Product-limit estimators of the survival function with twice censored data. *The Annals of Statistics*, **34(2)**, 925-938.

Robins, J. M. (1993), Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In: Proceedings of the American Statistical Association-Biopharmaceutical Section. Alexandria, Virginia: American Statistical Assocation, pp. 24-33.

Robins, J. M.;Finkelstein, D. (2000), Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted(IPCW) log-rank tests. *Biometrices*, **56**, 779-788.

Satten, G. A. and Datta S. (2001), The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, **55**, 207-210.

Shen, P-S. (2003), The product-limit estimate as an inverse-probability-weighted average.

Communications in Statisics, Theory and Methods, *32*, 1119-1133.

Tsai, W. Y. and Crowley, J. (1985), A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *The Annals of Statistics*, **13**, 1317-1334.

Turnbull, B. W. (1974), Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, **69**, 169-173.

Wang, M.-C. (1987), Product-limit estimates: a generalized maximum likelihood study. *Communications in Statistics, Theory and Methods*, **6**, 3117-3132.

Woodroofe, M. (1985), Estimating a distribution function with truncated data. *The Annals of Statistics*, **13**, 163-167.