Nonparametric Statistics, 2003, Vol. 15(2), pp. 253–265

Taylor & Francis
Taylor & Francis Group

NONPARAMETRIC ESTIMATION FOR MIDDLE-CENSORED DATA

S. RAO JAMMALAMADAKA[a,*] and VASUDEVAN MANGALAM[b]

[a]*Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA;* [b]*Department of Mathematics, Universiti Brunei Darussalam, Brunei*

(Received 2 November 1999; In final form 22 August 2000)

This paper provides the self-consistent estimator (SCE) and the nonparametric maximum likelihood estimator (NPMLE) for "middle-censored" data, in which a data value becomes unobservable if it falls within a random interval. We provide an algorithm to find the SCE and show that the NPMLE satisfies the self-consistency equation. We find a sufficient condition for the SCE to be concentrated on the uncensored observations. In addition, we find sufficient conditions for the consistency of the SCE and prove that consistency holds for the special case when one of the ends is a constant. Some simulation results and an illustrative example, using Danish melanoma data set, are provided.

Keywords: Survival function; Middle-censoring; Self-consistency; Nonparametric maximum likelihood estimation

AMS 1991 Subject Classifications: Primary: 62G05, 62G30; Secondary: 62G99

1 INTRODUCTION

Estimation of the unknown distribution of a random variable is of fundamental importance in statistics. In areas such as reliability, biometry, general medical follow-up studies and clinical trials, the distribution function of the underlying lifetime or more specifically. the survival function is of paramount interest.

In these situations, the random variable of interest is the lifetime and the observations refer to times of occurrence of an event such as death due to a certain cause under study, or times for equipment failure. When complete data are available, the Empirical Distribution Function (EDF) is used and it has many desirable properties. However, in many practical situations, it is quite common to have incomplete data, making the standard empirical distribution function (EDF) unavailable. Often, such incomplete observation of the data results from a random censoring mechanism. When observations are censored to the right, the product limit estimator due to Kaplan and Meier (1958) is used in place of the EDF and similar estimators exist for the left-censored case. Gehan (1965) and Turnbull (1974) and others considered doubly-censored data (where both left and right censoring occur simultaneously) and estimators for the distribution function have been developed. Groeneboom and Wellner

* Corresponding author.

ISSN 1048-5252 print; ISSN 1029-0311 online
DOI: 10.1080/1048525031000089266

(1992) and Geskus and Groeneboom (1996) studied the case of "interval-censored" data where one can only observe a censoring event and whether the time of the event of interest, say death, occurred before or after the occurrence of the censoring event. Nonparametric Maximum Likelihood Estimators (NPMLE) for the distribution of interest have been studied by various authors for all these cases. A Self Consistent Estimator (SCE) is usually obtained by solving a set of equations called the self-consistency equations (see Efron, 1967; Tarpey and Flury, 1996), and under some conditions this coincides with the NPMLE. Tsai and Crowley (1985) have shown that many of these cases can be unified by applying a generalized maximum likelihood principle. They also point out that solving the self-consistency equation is essentially equivalent to applying the EM algorithm for the corresponding missing data problem. See Dempster and Laird (1977) and McLachlan and Krishnan (1997) on the EM algorithm.

In this paper we consider an important variation and generalization of censoring where a data point becomes unobservable if it fell inside a random interval. When that happens we observe a censorship indicator and the interval of censorship. We will refer to this as "middle-censoring". Left censoring, right censoring and double censoring are special cases of this "middle censorship" by suitable choice of this censoring interval, which can be infinite. Middle-censoring where a random middle part is missing appears at first glance, as complementary to the idea of double-censoring where the middle is what is actually observed. However, if one considers these two schemes carefully along with the resulting data sets (see next section), they turn out to be quite distinct ideas.

Before discussing the estimator, we consider some situations where this type of censoring may arise. In general, in any lifetime study, if the subject is temporarily withdrawn from the study we will have an interval of censorship. It can be equipment failure that could occur during a period where observation is not possible or is not being made. In biomedical studies, the patient under observation may be absent from study for a short period during which time the event of interest may occur. As an example of double censoring, Turnbull (1974) refers to a study of African infant precocity by Leiderman *et al.* (1973), where establishing the for infant development for a community in Kenya was the purpose. A sample of 65 children are considered and each child was tested monthly to see if (s)he had learned to accomplish certain tasks. The time from birth to the learning time was the variable of interest. In their analysis, double-censoring occurred due to late arrivals (the child had already learned the skills before entering the study) and losses (the child had not acquired the skill by the end of time study). We envisage a scenario where there are no late entries or losses as such, but during a fixed time interval (this fixed interval is indeed, a random interval relative to the individual's lifetime) the observation was not possible, such as the temporary closure of the clinic due to an outbreak of say, war. If some children, of varying ages, developed the skill during this time, we do not observe the exact age of these children at the time of skill development, rather only the information that the event of interest occurred during a certain time interval. These ideas can, of course, be extended to more general random sets of censorship such as union of intervals or even more complicated sets but we have not explored this in detail.

In Section 2, we derive the self-consistency equation for the middle censored case and show that the NPMLE indeed satisfies the self-consistency equation. A simple example which shows how one computes the NPMLE is also given. In Section 3 we explore conditions under which the self-consistent estimator (SCE) is consistent and prove the consistency in an important special case. Section 4 illustrates the SCE for middle-censored case for a simulated data set as well as for a real data set on Melanoma survival, from Andersen *et al.* (1993). A computer program which allows the computation of the SCE is available by writing to the authors.

2 SELF-CONSISTENCY AND THE NPMLE

Let X_i, $i = 1, \ldots, n$, be a sequence of independent identically distributed (i.i.d.) random variables with unknown distribution F_0. Let $Y_i = (L_i, R_i)$ be a sequence of i.i.d. random vectors, independent of X_t's, with unknown bivariate distribution G such that $P(L_i < R_i) = 1$. While X denotes the variable of interest, Y represents the censoring mechanism. Using the notation

$$\delta_i = I[X_i \notin (L_i, R_i)],$$

we observe

$$\begin{aligned} Z_i = X_i \quad &\text{when } \delta_i = 1 \ \textit{ i.e.}, \text{ if } X_i \notin (L_i, R_i) \\ = (L_i, R_i) \quad &\text{when } \delta_i = 0 \ \textit{ i.e.}, \text{ if } X_i \in (L_i, R_i) \end{aligned}$$

That is, we either observe the original value X_i, if there is no censoring or the interval of censoring (L_i, R_i) when there is censoring.

In many censoring situations, if we were to try to estimate the distribution function via the EM algorithm the resulting equation takes the form

$$\hat{F}(t) = E_{\hat{F}}[E_n(t)|\mathbb{Z}]$$

as described by Tsai and Crowley (1985), where E_n is the empirical distribution function. This equation was first introduced and referred to as self-consistency equation by Efron (1967). In the middle censored case the SCE F_n, satisfies the equation

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i I(X_i \le t) + \bar{\delta}_i I(R_i \le t) + \bar{\delta}_i I[t \in (L_i,R_i)]\frac{F_n(t) - F_n(L_i)}{F_n(R_i-) - F_n(L_i)}\right\} \qquad (1)$$

where $\bar{\delta}_i = 1 - \delta_i$. (For the rest of the paper we will follow the convention that $\bar{x}$, for any variable or function x, indicates $1 - x$). As in the case of doubly-censored data, there is no explicit closed form solution to the equation and has to be computed by the iterative formula

$$\hat{F}^{(m+1)}(t) = E_{\hat{F}^{(m)}}[E_n(t)|\mathbb{Z}].$$

The convergence of the algorithm is assured by Theorem 2.1 of Tsai and Crowley (1985) provided that the initial estimator gives positive mass to all observed points. See Remark 2.1 below regarding the choice of the initial estimator. For a general discussion on self-consistency and its relation to EM algorithm, see Tarpey and Flury (1996).

Let $\mathcal{F}$ denote the set of all distribution functions on the line. For $F \in \mathcal{F}$ the likelihood of the sample is given by

$$L(F) = \prod_{i=1}^{n}[F(X_i) - F(X_i-)]^{\delta_i}[F(R_i-) - F(L_i)]^{1-\delta_i}.$$

Denoting by $\Delta F(x) = F(x) - F(x-)$, $\phi(F) \equiv (1/n)\log L(F)$ is given by

$$\begin{aligned} \phi(F) &= \frac{1}{n}\sum_{i=1}^{n}[\delta_i \log(\Delta F(X_i)) + \bar{\delta}_i \log[F(R_i-) - F(L_i)]] \\ &= \int \{I[x \notin (l, r)]\log \Delta F(x) + I[x \in (l, r)]\log[F(r-) - F(l)]\}\, dP_n(x, l, r) \end{aligned}$$

where P_n is the empirical measure of $\{(X_i, L_i, R_i): 1 \le i \le n\}$. The maximizer of ϕ is clearly the NPMLE. In the next theorem, we show that the NPMLE for middle censored data satisfies the self-consistency equation. But before that we need the following lemma.

LEMMA 1 *Define*

$$\begin{aligned} A_t(x) &= \frac{F(t \wedge x)}{F(t)} - F(x) \quad \text{if } F(t) > 0 \\ &= 0 \qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{2}$$

where $t \wedge x = \min(t, x)$. *Then*

$$K(x) = F(x) + hA_t(x)$$

defines a class of distribution functions for h sufficiently close to zero.

Proof Note that we need to show this only for $F(t) > 0$ since $A_t \equiv 0$ when $F(t) = 0$.

$$K(x) = (1 - h)F(x) + h\frac{F(t \wedge x)}{F(t)}$$

is a convex combination of two cdf's and hence is a cdf for $0 \le h < 1$. For negative h, write $K = F - hA_t$ with $h > 0$ so that

$$K(x) = (1 + h)F(x) - h\frac{F(t \wedge x)}{F(t)}.$$

Clearly $K(-\infty) = 0$ and $K(\infty) = 1$. It is also right-continuous so it remains to show that it is monotone. We check this separately on $(-\infty, t]$ and $[t, \infty)$. For x in $(-\infty, t]$,

$$\begin{aligned} K(x) &= (1 + h)F(x) - h\frac{F(x)}{F(t)} \\ &= F(x)\left(1 - h\frac{(1 - F(t))}{F(t)}\right). \end{aligned}$$

This is clearly bounded by 1 and is non-negative if $h \le (F(t)/1 - F(t))$ and in this case, K is monotone non-decreasing. Similarly on $[t, \infty)$,

$$\begin{aligned} K(x) &= (1 + h)F(x) - h \\ &= F(x) - h(1 - F(x)). \end{aligned}$$

Again this is bounded by 1 and is non-negative so long as $h \le (F(x)/1 - F(x))$ which is assured if $h \le (F(t)/1 - F(t))$ since $t \le x$. Monotonicity of K is clear. ■

THEOREM 1 *The NPMLE satisfies the equation*

$$F(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i I(X_i \le t) + \bar{\delta}_i I(R_i \le t) + \bar{\delta}_i I[t \in (L_i, R_i)]\frac{F(t) - F(L_i)}{F(R_i-) - F(L_i)}\right\}. \tag{3}$$

Proof If F maximizes ϕ, then the directional derivative of ϕ towards A_t should be zero at F *i.e.*, satisfies the equation

$$\begin{aligned}0 &= \lim_{h\to 0}\frac{\phi(F+hA_t)-\phi(F)}{h}\\ &= \int\Bigg\{I[x\notin(l,r)]\lim_{h\to 0}\frac{\log[\Delta F(x)+h\Delta A_t(x)]-\log\,\Delta F(x)}{h}\\ &\quad +I[x\in(l,r)]\lim_{h\to 0}\frac{\log[F(r-)-F(l)+h(A_t(r-)-A_t(l))]-\log[F(r-)-F(l)]}{h}\Bigg\}\\ &\quad \times \mathrm{d}P_n(x,l,r)\end{aligned}$$

as the integral involved is really a finite sum and hence interchanging of limit and integration is valid. When $F(t)>0$, the first of the two limits inside the integral is

$$\Delta A_t(x)\lim_{\varepsilon\to 0}\frac{\log[\Delta F(x)+\varepsilon]-\log\,\Delta F(x)}{\varepsilon}=\frac{\Delta A_t(x)}{\Delta F(x)}=\frac{I(x\le t)}{F(t)}-1.$$

The second limit is similarly equal to

$$\frac{A_t(r-)-A_t(l)}{F(r-)-F(l)}=\frac{F(t)\wedge F(r-)-F(t\wedge I)}{F(t)[F(r-)-F(l)]}-1$$

where $F(t)\wedge F(r-)$ stands for $F(t)$ if $t<r$ and $F(r-)$ otherwise. Thus we get

$$1=\int\left\{I[x\notin(l,r)]\frac{I(x\le t)}{F(t)}+I[x\in(l,r)]\frac{F(t)\wedge F(r-)-F(t\wedge I)}{F(t)[F(r-)-F(l)]}\right\}\mathrm{d}P_n(x,l,r)$$

or

$$\begin{aligned}F(t)=\int\Bigg\{&I[x\notin(l,r)]I(x\le t)+I(l<x<r\le t)+I[x,t\in(l,r)].\\ &\times\frac{F(t)-F(l)}{F(r-)-F(l)}\Bigg\}\mathrm{d}P_n(x,l,r).\end{aligned}$$

RHS of the above is same as RHS of (3). ■

It is a question of considerable interest to ask if NPMLE will have all its mass on the uncensored observations. The answer is yes, provided all censored intervals contain at least one uncensored observation. When there is a censoring interval empty of uncensored observations, clearly some mass has to be attached to that interval or the likelihood would be zero. That the weights are concentrated on the uncensored observations when all censoring intervals are non-empty is a consequence of the following proposition.

PROPOSITION 1 *If each observed censored interval* (L_i, R_i) *contains at least one uncensored observation* X_j, $j\neq i$, *then any distribution function that satisfies* (3) *attaches all its mass on the uncensored observations.*

Proof Let F be a distribution satisfying (3) and let $x_1, x_2, \ldots, x_m$ be the uncensored observations. For any x let $\Delta F(x) = F(x) - F(x-)$ be the weight F associates to x. We need to show that $\sum_{j=1}^{m} \Delta F(x_j) = 1$. From (3) it follows that

$$\Delta F(x_j) = \frac{1}{n} + \frac{1}{n}\sum_{i=1}^{n}\left\{(1-\delta_i) I[x_j \in (L_i, R_i)] \frac{\Delta F(x_j)}{F(R_i-) - F(L_i)}\right\}. \tag{4}$$

Summing (4) over all uncensored observations, we get

$$\sum_{j=1}^{m} \Delta F(x_j) = \frac{m}{n} + \frac{1}{n}\sum_{i=1}^{n} \frac{(1-\delta_i)\sum_j I[x_j \in (L_i, R_i)]\Delta F(x_j)}{F(R_i-) - F(L_i)} \tag{5}$$

For each censoring interval (L_i, R_i), let α_i be the slack between the mass associated to the interval and the sum of weights of all uncensored observations in the interval. Then

$$\alpha_i = F(R_i-) - F(L_i) - \sum_{j=1}^{m} I[x_j \in (L_i, R_i)]\Delta F(x_j) \tag{6}$$

and α_i's are all non-negative. From (5) and (6), it follows that

$$1 - \sum \alpha_i = 1 - \frac{1}{n}\sum \frac{\alpha_i}{F(R_i-) - F(L_i)}$$

or

$$\sum \alpha_i = \frac{1}{n}\sum \frac{\alpha_i}{F(R_i-) - F(L_i)} \tag{7}$$

where the sum is over all censored observations. As every interval contains at least one uncensored observation, it follows from (4) that $F(R_i-) - F(L_i) \geq (1/n)$ and hence (7) implies that

$$\alpha_i = \frac{\alpha_i}{n(F(R_i-) - F(L_i))} \tag{8}$$

for each i. Now if there exists i such that $\alpha_i > 0$, $(F(R_i-) - F(L_i))(1/n) + \alpha_i > (1/n)$ contradicting (8). ■

We have now proved that the NPMLE will have all its mass on the uncensored observations except when it so happens that a censored interval contains no uncensored observation. If this happens, we are in a situation similar to that of right censored data where the largest observation is censored. While in the right censored case the extra mass is usually left unassigned, for middle-censored data there is a natural way of handling this. When a censored interval contains no uncensored points, we let the mass that corresponds to that interval be assigned to its midpoint. Thus our initial estimator may give equal mass to all uncensored observations and to the midpoints of those finite censored intervals that contain no uncensored observations. If an infinite censoring interval happens to be empty of uncensored observations, one can then assign the mass to any arbitrary point inside this interval for the estimator to have a maximum.

Consider the following example where $n = 5$ and $z_1 = 2$, $z_2 = 4$, $z_3 = 6$, $z_4 = (1, 5)$ and $z_5 = (3, 7)$. Let p_1, p_2, p_3 be the masses to be assigned to z_1, z_2, z_3 respectively. The likelihood function is given by

$$p_1 p_2 p_3 (p_1 + p_2)(p_2 + p_3)$$

and, as p_i's add up to 1 and the roles of p_1 and p_3 are interchangeable, we can simplify the problem to that of maximizing $(x^2)(1-2x)(1-x)^2$ with $p_1 = p_3 = x$ and $p_2 = 1-2x$. The solution, then, is given by $x = (5-\sqrt{5})/10$ so that $p_1 = p_3 = (5-\sqrt{5})/10$ and $p_2 = 1/\sqrt{5}$ is the solution to the NPMLE. In this example the iterations of the self-consistency equation rapidly converged to the NPMLE.

The SCE, being a result of convergence of the EM algorithm, provides a local maximum of the likelihood equation [see, for example, Mykland and Ren, 1996] and may not coincide with the NPMLE. Examples of cases when an SCE is not the NPMLE can be constructed by considering situations where two empty censoring intervals overlap. For instance, if we have 1, 2, (3, 6), (4, 7) as the data, we could assign 0.25 mass to 1, 2, 4.5 and 5.5 to get an SCE. The NPMLE will assign 0.25 each on 1 and 2, but assign 0.5 on some point, say 5, on the overlap area (4, 6). Both estimators are self-consistent, but the latter has higher likelihood. This happens whenever there are empty, overlapping intervals. In the next section we shall show the strong consistency of self-consistent estimators for certain cases. This will demonstrate that SCE and NPMLE are, at least for these special cases, asymptotically equivalent and hence will be approximately the same for large samples.

3 CONSISTENCY OF SELF-CONSISTENT ESTIMATORS

Define P and Q, sub-distribution functions on $\mathbb{R}$ and $\mathbb{R}^2$ respectively, by

$$P(t) = P(X \le t, \delta = 1)$$
$$Q(l, r) = P(L \le l, R \le r, \delta = 0)$$

and their empirical versions P_n and Q_n by

$$P_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t, \delta_i = 1)$$
$$Q_n(l, r) = \frac{1}{n}\sum_{i=1}^{n} I(L_i \le l, R_i \le r, \delta_i = 0).$$

Then by Glivenko-Cantelli Lemma, it follows that P_n and Q_n converge almost surely to P and Q respectively and the convergence in each case is uniform on the respective domain. Also, (1) can be written in terms of P_n and Q_n as follows:

$$F_n(t) = P_n(t) + \int \frac{F_n(t) \wedge F_n(r-) - F_n(t \wedge l)}{F_n(r-) - F_n(l)}\, \mathrm{d}Q_n(l, r) \tag{9}$$

By Helly's Theorem, $\exists$ a subsequence n_k and a non-decreasing function F bounded by 0 and 1 such that on a set of probability 1, $F_{n_k}(t) \to F(t)$ for each t.

PROPOSITION 2 *If $\{\varphi_n\}$ is a sequence of functions on $\mathbb{R}^2$ which converge uniformly to bounded continuous function φ, then*

$$\int \varphi_n(l, r)\, \mathrm{d}Q_n(l, r) \to \int \varphi(l, r)\, \mathrm{d}Q(l, r).$$

Proof Note that,

$$\begin{aligned}
&\left|\int \varphi_n(l,r)\,\mathrm{d}Q_n(l,r) - \int \varphi_n(l,r)\,\mathrm{d}Q(l,r)\right| \\
&\quad\le \left|\int (\varphi_n(l,r)-\varphi(l,r))\,\mathrm{d}Q_n(l,r)\right| + \left|\int \varphi(l,r)\,\mathrm{d}Q_n(l,r) - \int \varphi(l,r)\,\mathrm{d}Q(l,r)\right| \\
&\quad\le \|\varphi_n-\varphi\| \int \mathrm{d}Q_n + \left|\int \varphi(l,r)\,\mathrm{d}Q_n(l,r) - \int \varphi(l,r)\,\mathrm{d}Q(l,r)\right|
\end{aligned}$$

where $\|\cdot\|$ represents the supremum norm. Now the first term on the RHS of the inequality goes to zero since $\int \mathrm{d}Q_n = 1$ while the second term goes to zero because the sequence of empirical measures Q_n converge to Q weakly and φ is a bounded continuous function. ■

LEMMA 2 *Any subsequential limit F of F_n will satisfy the equation*

$$F(t) = P(t) + \int \frac{F(t)\wedge F(r-) - F(t\wedge l)}{F(r-)-F(l)}\,\mathrm{d}Q(l,r) \tag{10}$$

Proof For a fixed t, taking limits in (9) through the subsequence n_k as $k\to\infty$ and using Proposition 2 with

$$\varphi_n(l,r) = \frac{F_n(t)\wedge F_n(r-) - F_n(t\wedge l)}{F_n(r-)-F_n(l)},$$

the result follows. ■

When P and Q are written in terms of F_0 and G, (10) is equivalent to

$$F(t)-F_0(t) = \int_{l<t<r}\left[\frac{F(t)-F(l)}{F(r-)-F(l)}(F_0(r)-F_0(l)) + F_0(l) - F_0(t)\right]\mathrm{d}G(l,r). \tag{11}$$

From (11), it follows that $F(\infty)=1$ and $F(-\infty)=0$. Note that if $F=F_0$, (11) is automatically satisfied. If we were able to show that (11) has a unique solution, then it follows that F_0 is the only limit point of $\{F_n\}$. Then we will have that on a set of probability 1, $F_n(x)\to F_0(x)$ for each x and by continuity of F_0 uniformity of the almost sure convergence follows.

A necessary condition for consistency is what we call "identifiability". Let $A(t) = P(L<t<R)$. The condition is that A be not identically 1 on any interval $[a,b]$, $a\le b$ for which $F_0(b) > F_0(a-)$. Observe that if $A\equiv 1$ on any interval where F_0 has a positive mass, then censoring occurs with probability 1 on such an interval. As a consequence, there will be no observations on this interval and that prevents us from distinguishing any two distributions which are identical outside $[a,b]$ but differing only on $[a,b]$. This condition will be referred to as the "identifiability condition" and is a requirement for consistent estimation of F_0.

LEMMA 3 *Let $h = (F-F_0)$ and*

$$g(t) = \int_{\mathbb{R}^2} \psi(l,r) I(l<t<r)\,\mathrm{d}G(l,r)$$

where $\psi(l, r) = (h(r) - h(l)/F(r-) - F(l))$. *Then*

$$\bar{A}\mathrm{d}h = -g\mathrm{d}F \tag{12}$$

Proof From (11) we get

$$\begin{aligned} h(t) &= -\int_{l<t<r} \left\{ (F(t) - F(l)) \frac{h(r) - h(l)}{F(r-) - F(l)} + h(l) - h(t) \right\} \mathrm{d}G(l, r) \\ &= -\int_{l<t<r} [(F(t) - F(l))\psi(l, r) + h(l)]\mathrm{d}G(l, r) + h(t)A(t). \end{aligned}$$

So,

$$\begin{aligned} -h(t)\bar{A}(t) &= \int_{l<t<r} [(F(t) - F(l))\psi(l, r) + h(l)]\mathrm{d}G(l, r) \\ &= F(t)g(t) + C(t) \end{aligned} \tag{13}$$

where

$$C(t) = \int_{l<t<r} [h(l) - F(l)\psi(l, r)]\mathrm{d}G(l, r).$$

"Differentiating" both sides of (13) w.r.t. t, we get

$$h(t)\mathrm{d}A(t) - \bar{A}(t)\mathrm{d}h(t) = g(t)\mathrm{d}F(t) + F(t)\mathrm{d}g(t) + \mathrm{d}C(t).$$

To show that (12) holds, clearly it is sufficient to show that

$$F(t)\mathrm{d}g(t) + \mathrm{d}C(t) - h(t)\mathrm{d}A(t) = 0 \tag{14}$$

If $B(t) = \int_{l<t<r} H(l, r)\,\mathrm{d}G(l, r)$ for some function H, then it can be shown that $\mathrm{d}B(t) = (\int_t^\infty H(t, r)\,\mathrm{d}F_{R|L}(r|t))\mathrm{d}F_L(t) - (\int_{-\infty}^t H(l, t)\,\mathrm{d}F_{L|R}(l|t))\mathrm{d}F_R(t)$ where $F_{R|L}(\cdot\,|t)$ is the conditional distribution function of R given $L = t$. Hence, applying this to g and C,

$$\begin{aligned} \text{LHS of}(14) &= F(t)\left[\int_t^\infty \psi(t, r)\,\mathrm{d}F_{R|L}(r|t)\right]\mathrm{d}F_L(t) - F(t)\left[\int_{-\infty}^t \psi(l, t)\,\mathrm{d}F_{L|R}(l|t)\right]\mathrm{d}F_R(t) \\ &\quad + \left[\int_t^\infty (h(t) - F(t)\psi(t, r))\mathrm{d}F_{R|L}(r|t)\right]\mathrm{d}F_L(t) \\ &\quad - \left[\int_{-\infty}^t (h(l) - F(l)\psi(l, t))\mathrm{d}F_{L|R}(l|t)\right]\mathrm{d}F_R(t) - h(t)\,\mathrm{d}A(t) \\ &= -\mathrm{d}F_R(t)\int_{-\infty}^t h(t)\,\mathrm{d}F_{L|R}(l|t) + \mathrm{d}F_L(t)\int_t^\infty h(t)\,\mathrm{d}F_{R|L}(r|t) - h(t)\,\mathrm{d}A(t) \\ &= h(t)(\mathrm{d}F_L(t) - \mathrm{d}F_R(t)) - h(t)\,\mathrm{d}A(t) \\ &= 0 \end{aligned}$$

because $\int_{-\infty}^t \mathrm{d}F_{L|R}(l|t) = \int_t^\infty \mathrm{d}F_{R|L}(r|t) = 1$ and $A(t) = F_L(t) - F_R(t)$. ■

Thus, if the only function h satisfying (12) is the zero function, then we would have proved the strong consistency of the SCE. We have not yet been able to show that this (12) has a unique solution in the general case, but we give below a proof for the special case when one of the end points of the censoring interval is degenerate. Although the result is stated for the case when L is degenerate (for instance, censoring if it occurs, starts on a certain birthday of the individual), the proof works equally well when R is degenerate.

THEOREM 2 *Assume that F_0 and F_R are continuous and $L \equiv l_0$. Assume the identifiability condition is satisfied. Then the only function F that satisfies* (11) *is F_0 and hence the SCE is uniformly strongly consistent.*

Proof In this special case (13) becomes

$$-h(t)\bar{A}(t) = I(t > l_0)\int_t^\infty [(F(t) - F(l_0))\psi(l_0, r) + h(l_0)]\mathrm{d}F_R(r) \tag{15}$$

As $\bar{A}(t) = 1 - P[t \in (l_0, R)] = 1 - I(t > l_0)\bar{F}_R(t)$, $\bar{A}(t) = 1$ for all $t \le l_0$ and $\bar{A}(t) = F_R(t)$ for all $t > l_0$; hence from (15), $h(t) = 0$ for all $t \le l_0$. In particular, $h(l_0) = 0$. Thus (15) becomes

$$-h(t)\bar{A}(t) = I(t > l_0)\int_t^\infty \frac{F(t) - F(l_0)}{F(r-) - F(l_0)} h(r)\,\mathrm{d}F_R(r).$$

Similarly, (12) holds with

$$g(t) = I(t > l_0)\int_t^\infty \frac{h(r)}{F(r-) - F(l_0)}\,\mathrm{d}F_R(r). \tag{16}$$

Note that from the assumptions of the theorem it follows that F, h and g are continuous on (l_0, ∞). We aim to show that $h \equiv 0$ on (l_0, ∞). Assuming $\exists t_0 > l_0$ such that $h(t_0) > 0$, we will arrive at a contradiction. The proof is similar if $h(t_0) < 0$. ■

As $h(l_0) = h(\infty) = 0$, $\exists t_1 < t_2$ such that $t_1 \ge l_0$, $t_2 \le \infty$, $h(t_1) = h(t_2) = 0$ and $h(t) > 0$ on (t_1, t_2). From (16), on (l_0, ∞), $g(t) = \int_t^\infty \psi(r)\,\mathrm{d}F_R(r)$ where $\psi(r) = (h(r)/F(r-) - F(l_0))$.

CLAIM 1 $g(t_1) \le 0$.

Suppose not. Then $\exists t^*$ such that $g > 0$ on (t_1, t^*). We shall now show that $\bar{A}(t) > 0$ on (t_1, t^*). If $\exists \tau \in (t_1, t^*)$ such that $\bar{A}(\tau) = 0$, then $\bar{A}(t) = 0$ for all $t \in (l_0, \tau)$ so that by the identifiability condition, $\mathrm{d}F_0(t) = 0$ for all $t \in (l_0, \tau)$. From (12) $\mathrm{d}F(t) = 0$ on (t_1, τ); so $\mathrm{d}h(t) = \mathrm{d}F(t) - \mathrm{d}F_0(t) = 0$ on (t_1, τ) which implies $h \equiv 0$ there, contrary to our assumption.

From (12), $\mathrm{d}h(t) = (-g(t)/1 - A(t))\mathrm{d}F(t)$, so $\int_{t_1}^{t^*} (-g(t)/1 - A(t))\mathrm{d}F(t) = h(t^*) - h(t_1) = h(t^*)$. Now, LHS ≤ 0 contradicting the fact that $h(t^*) > 0$. This proves Claim 1.

CLAIM 2 $g(t_2) \ge 0$.

Suppose not. Then $\exists t^* \in (t_0, t_2)$ such that $g < 0$ on (t^*, t_2). Similar to the previous situation, we have $\bar{A}(t) > 0$ on (t^*, t_2). As earlier, $h(t_2) - h(t^*) = \int_{t^*}^{t_2} (-g(t)/1 - A(t))\mathrm{d}F(t) \ge 0$, implying $h(t^*) \le 0$ which is contradiction. Thus Claim 2 is proved.

On (t_1, t_2), $\mathrm{d}g(t) = -\psi(t)\mathrm{d}F_R(t) = (-h(t)\mathrm{d}F_R(t)/F(t-) - F(l_0)) \le 0$, so g is decreasing there. Thus from Claim 1 and Claim 2 it follows that $g \equiv 0$ on (t_1, t_2). (Note that the argument goes through even if $t_2 = \infty$). From (12), $\bar{A}\mathrm{d}h \equiv 0$ on (t_1, t_2). As $g(t) = \int_t^\infty \psi(r)\,\mathrm{d}F_R(r)$, F_R is a constant (t_1, t_2) and hence $\bar{A}$ is a constant on (t_1, t_2). If $c > 0$, $h \equiv 0$ on (t_1, t_2) which is a contradiction. If $c = 0$, $\bar{A} \equiv 0$ on (t_1, t_2) and hence by the identifiability condition F_0 is constant on (t_1, t_2). As $h(t_1) = h(t_2) = 0$, $F(t_1) = F(t_2)$, so F is a constant on (t_1, t_2). So h is a constant on (t_1, t_2), which means $h \equiv 0$ on (t_1, t_2). This is a contradiction.

4 ILLUSTRATIVE EXAMPLES

A simulation study was performed to measure the performance of self-consistent estimators where an exponential (mean = 10) random variable was middle-censored by random intervals

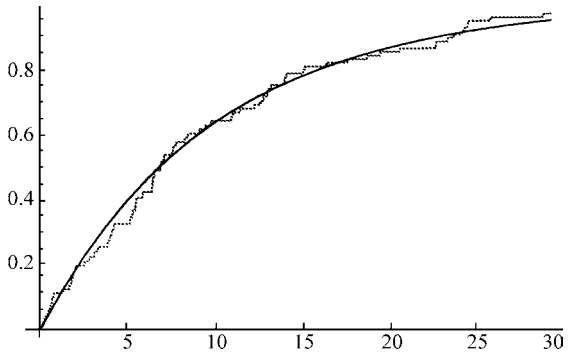

FIGURE 1 The EDF and SCE for the simulated exponential data.

with left end points being exponential (mean = 5) and interval widths being an independent exponential (mean = 5). A sample of size 100 was taken and this resulted in 22 of them being censored. Figure 1 shows the SCE along with the original exponential distribution function F_0. The maximum distance $\|F_n - F_0\|$ is 0.0827 which is very good compared to the Kolmogorov–Smirnov distance, namely the maximum distance of the EDF. E_n of the uncensored data from the true distribution, which is $\|E_n - F_0\| = 0.0715$. The authors also tried out various other survival distributions such as gamma and Weibull that were censored by intervals whose left ends were distributed as exponential, gamma, Weibull or uniform and interval width was a positive random variable or a constant. In all these cases, the resulting estimators for middle censoring were in very close agreement with the EDF of the original uncensored data. It is clear that the amount of censoring in any of these cases, is approximately $P(L \leq X \leq R)$.

Finally we applied our techniques to an actual data set on melanoma survival collected at Odense University Hospital, Denmark [see Andersen *et al.*, 1993]. The sample contains 205 data points, ranging from 10 to 5565. The data were censored by a random interval whose left end was an exponential random variable with mean 2000 and width was exponential with mean 1000. Over 23% of data were censored resulting in 157 uncensored observations. The SCE F_n is given in Figure 2 while the EDF E_n of the survival data is in Figure 3.

They are shown super-imposed in Figure 4, to see how close they are. Indeed, the maximum distance $\|F_n - E_n\|$ between them is 0.0604 while the maximum relative distance $\|((F_n - E_n)/E_n)\|$ turns out to be still a small 0.153.

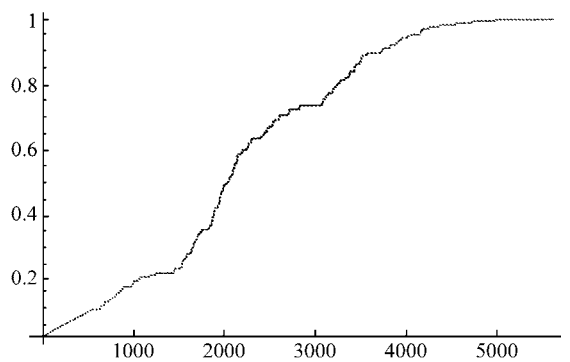

FIGURE 2 SCE for the censored melanoma survival data.

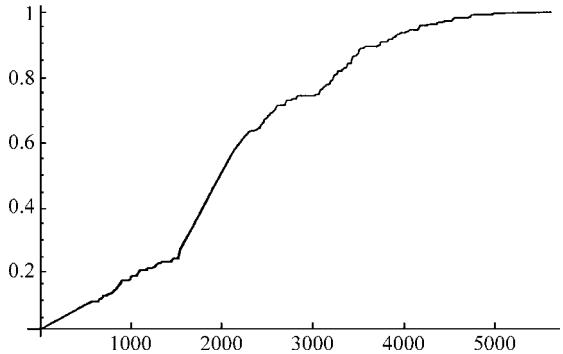

FIGURE 3 EDF for the uncensored melanoma survival data.

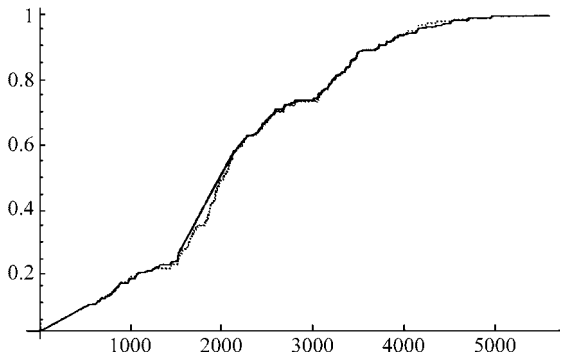

FIGURE 4 EDF and SCE superimposed.

Acknowledgements

We would like to thank an anonymous referee whose persistence led to a much more readable paper.

References

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser.*, **B39**, 1–38 (with discussion).

Efron, B. (1967). The two-sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Stat. Probab.*, Vol. 4. University of California Press, Berkeley, pp. 831–853.

Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*, **52**, 650–653.

Geskus, R. B. and Groeneboom, P. (1996). Asymptotically optimal estimation of smooth functionals for interval censoring. *J. Statist. Neerlandica*, **50**, 69–88.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Non-parametric Maximum Likelihood Estimation.* DMV Seminar, Vol. 19. Birkhäuser Verlag, Basel.

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

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **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.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley and Sons, Inc., New York.

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

Tarpey, T. and Flury, B. (1996). Self-consistency: A fundamental concept in statistics. *Statistical Science*, **11**, 229–243.

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

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, **69**, 169–173.