

Approximate Self Consistency for Middle-Censored Data

by

S. Rao Jammalamadaka
Department of Statistics & Applied Probability,
University of California, Santa Barbara, CA 93106, USA.

and

Srikanth K. Iyer
Department of Mathematics, Indian Institute of Technology
Kanpur, 208016, India.

Abstract

Middle censoring refers to data that becomes unobservable if it falls within a random interval. The lifetime distribution of such data is defined via the self-consistency equation. We propose an approximation to this distribution function for which an estimator and its asymptotic properties are very easy to establish.

Abbreviated title: Estimation for middle-censored data.

Keywords: Middle censoring, self-consistent estimator, consistency, weak convergence.

AMS 1991 subject classifications: Primary 62G05, 62G30; Secondary 62G99.

1 Introduction

1.1 Middle Censoring

Our aim in this paper is the estimation of the lifetime distribution or its complement, the survival distribution, for middle-censored data. Middle censoring occurs when a data point becomes unobservable if it falls inside a random interval. This is a generalization of left and right censored data and is quite distinct from the case of doubly censored data.

In situations where data is not subject to any censoring it is natural to use the empirical distribution function (EDF) to estimate the lifetime distribution of the population. If the data is subject to censoring, the EDF is unavailable and modifications have to be made to the EDF to estimate the lifetime distribution. In such a case, the lifetime distribution is constructed based on the available information, that is the lifetimes of the individuals that do not fall in the censoring interval as well as the censoring information. This reconstruction is done using the self consistency equation (see eg. Tarpey and Flury (1996)) and forms the basis for defining self consistent estimators (SCE) in place of the unavailable EDF, to estimate lifetimes.

In case of right censored data, the well known product limit estimator due to Kaplan and Meier (1958) is used and similar estimates exist for the left censored case. Gehan (1965), Turnbull (1974) and others consider doubly censored data (where both left and right censoring can occur simultaneously), while Groeneboom and Wellner (1992) and Geskus and Groeneboom (1996) study the case of interval censored data. Non-parametric maximum likelihood estimators (NPMLE) and SCE have been obtained for the above cases and these coincide under certain conditions. Tsai and Crowley (1985) have shown that many of these cases can be unified by a generalized maximum likelihood principle. It is pointed out in that paper that solving for a self consistent estimator is akin to applying the EM algorithm. The idea of middle-censoring was introduced and an NPMLE obtained for such data, by Jammalamadaka and Mangalam (2000), hereafter referred to as JM. This paper also illustrates via concrete real data situations where middle-censoring is applicable. JM showed that the NPMLE is a SCE, but consistency was proved for the SCE under the rather stringent condition that one of the ends of the censoring intervals is a constant.

1.2 Censoring and self consistency

In case of censored data one typically looks at an estimator that satisfies a self consistency equation. Often it is not possible to obtain a closed form solution for this equation and hence the estimator has to be computed using iterative methods as is

done in JM. Further it is proved there that under certain conditions this estimator converges to the solution of an equation which is the lifetime distribution of the population. We look at this problem in a slightly different way and address the practical difficulties that arise in the use of SCE. We suggest a simpler alternative estimator which is not recursive and for which not only consistency but weak convergence can be established.

Suppose the lifetimes denoted by X follow an unknown distribution F_0 , and our goal is to estimate this F_0 . Corresponding to every individual in the population there is a censoring interval distributed as the random interval (L, R) , independent of the lifetime, with unknown bivariate distribution G . During the time period (L, R) no observation is possible. That is, for any individual with lifetime X , let $\delta := I[X \notin (L, R)]$. If $\delta = 1$ then we can observe X , else we can observe only the censored interval (L, R) corresponding to this individual. Thus some information regarding lifetimes is missing in the sample. So we reconstruct this information based on the uncensored lifetimes and the self consistency equation.

Let Z represent the observable, i.e.,

$$Z = \begin{cases} X & \text{if } X \notin (L, R) \text{ (i.e. } \delta = 1) \\ (L, R] & \text{otherwise.} \end{cases} \quad (1.1)$$

Define P and Q , subdistributions on \mathfrak{R} and \mathfrak{R}^2 respectively as

$$\begin{aligned} P(t) &= P(X \leq t, \delta = 1) \\ Q(l, r) &= P(L \leq l, R \leq r, \delta = 0). \end{aligned} \quad (1.2)$$

P is the lifetime distribution of the uncensored observation and Q governs the distribution of censoring intervals of the censored observations. Note that Q is concentrated on the region $l \leq r$. We shall make the assumption that both P and Q are continuous.

The population lifetime distribution is defined via the solution of the self consistency equation (see JM)

$$F(t) = P(t) + \int \frac{F(t \wedge r) - F(t \wedge l)}{F(r) - F(l)} dQ(l, r). \quad (1.3)$$

In order to understand why the lifetime distribution should satisfy this equation, let us rewrite (1.3) as follows,

$$\begin{aligned} F(t) &= P(t) + \int_{l \leq r \leq t} dQ(l, r) + \int_{l < t < r} \frac{F(t) - F(l)}{F(r) - F(l)} dQ(l, r) \\ &= P(t) + Q(t, t) + \int_{l < t < r} \frac{F(t) - F(l)}{F(r) - F(l)} dQ(l, r) \end{aligned} \quad (1.4)$$

The first two terms represent the uncensored individuals who have lifetimes less than t and the censored individuals for whom the censoring interval lies in $(-\infty, t]$. What is not available at time t is the information regarding the lifetimes of censored individuals whose censoring interval contains t . This is given by the third term where the integrand is the conditional probability of the lifetime taking values in (l, t) given that it is in (l, r) . Thus it makes sense to define the lifetime distribution according to (1.3). Note that F_0 satisfies (1.3). Once the population lifetime distribution is specified, it is easy to define self consistent estimators.

Given data $\{Z_i, i = 1, \dots, n\}$, we define the empirical versions of P, Q and F as

$$\begin{aligned} P_n(t) &= \frac{1}{n} \sum_{i=1}^n I(X_i \leq t, \delta_i = 1) \\ Q_n(l, r) &= \frac{1}{n} \sum_{i=1}^n I(L_i \leq l, R_i \leq r, \delta_i = 0) \end{aligned} \quad (1.5)$$

$$F_n(t) = P_n(t) + \int \frac{F_n(t \wedge r) - F_n(t \wedge l)}{F_n(r) - F_n(l)} dQ_n(l, r) \quad (1.6)$$

Since (1.6) does not have a closed form solution, the SCE must be computed using iterative methods.

The Identifiability Condition. Consistency of estimators in case of censored data is usually proved under the so called identifiability condition. Let $A(t) = P(L < t < R)$. For any $a \leq b$ for which $A(t)$ is identically 1 on $[a, b]$, $F_0(b) = F_0(a)$. If censoring occurs on an interval with probability 1, then it is impossible to estimate F_0 consistently in that interval.

Remark. The main difficulty in establishing the consistency of the SCE lies in showing that there is a unique solution to (1.3) given P and Q . JM (see section 3) prove consistency under the rather restrictive condition that one of the end points of the censoring interval is a constant. The question of weak convergence is even more daunting (see eg. Chang (1990) in double-censoring context) for our original estimator. As it happens in other areas of statistics, when a problem is very difficult to solve, we take easier ways out by considering alternative procedures with comparable efficiency. Our alternate estimator is offered in that spirit.

2 Approximate self consistency and main results

In order to get around the difficulties mentioned above, we propose a modification to the SCE of the lifetime distribution proposed in (1.6). If one examines the distribution given by (1.4) or, more simply, if not so precisely (1.6), what is happening is that

the mass of the censored data is being reallocated to the uncensored observations in a meaningful (self-consistent) way.

The question we ask is whether we can make this reallocation in a simple yet reasonable way so that the problems of computing the estimator and the asymptotics for the estimator becomes straightforward. To this end we define an approximately self-consistent lifetime distribution based on P and Q . This is obtained by modifying (1.4) by

$$F^1(t) := P(t) + Q(t, t) + \int_{l < t < r} \frac{P(t) - P(l)}{P(r) - P(l)} dQ(l, r). \quad (2.1)$$

What this does is to reallocate the unexplained mass $1 - P(\infty)$ to the uncensored observations. This is done by assuming that a censored individual with censoring interval (L, R) will have a lifetime that is equally likely to be one of the uncensored lifetimes to fall in (L, R) . Note that under the identifiability condition F^1 is a proper distribution function.

A natural estimator for the above distribution would be

$$F_n^1(t) := P_n(t) + Q_n(t, t) + \int_{l < t < r} \frac{P_n(t) - P_n(l)}{P_n(r) - P_n(l)} dQ_n(l, r). \quad (2.2)$$

Remark. The integrand in (2.2) is taken to be zero if for any $t \in (L_i, R_i]$ for which $\delta_i = 0$, and the interval $(L_i, R_i]$ does not contain any uncensored observation. For all such intervals, F_n^1 puts mass $1/n$ at $t = R_i$. This will accrue due to the second term that will have a jump of size $1/n$ at $t = R_i$.

Remark. The above estimator can be easily extended to the case of left and right censored data. This is the case when the $Q(\ell, r)$ puts positive mass at $\ell = 0$, and $r = \infty$. The only modification one needs to make in the above estimator is to cover cases when an observation that is right censored, and the censoring interval does not contain any uncensored observations. In this case, we add a term $\epsilon_n(t)$ to $F_n^1(t)$, where ϵ_n puts mass $1/n$ on all L_i for which $R_i = \infty$ and $\delta_i = 0$ and the interval (L_i, ∞) does not contain any uncensored observation.

$$\epsilon_n(t) = \frac{1}{n} \sum_{i=1}^n I(L_i \leq t, R_i = \infty, \delta_i = 0, X_j \notin (L_i, \infty) \forall j : \delta_j = 1, j = 1, \dots, n) \quad (2.3)$$

Observe that under the identifiability condition ϵ_n will converge to zero a.s. uniformly in t by the Glivenko-Cantelli lemma. Thus, the asymptotic results given below will remain unaltered in the presence of right and left censoring.

Main Results

Suppose the identifiability condition given in Section 1.2 is satisfied and the functions P and Q in (2.1) are continuous. Then the results of this section establish the consistency and weak convergence of our estimator.

Theorem 2.1 (Consistency) *As $n \rightarrow \infty$, F_n^1 converges a.s. to F^1 uniformly in t .*

To begin with, we note that P_n and Q_n will converge to P and Q uniformly as a consequence of the Glivenko-Cantelli Lemma. The following lemma will be used in the proof of Theorem 2.1.

Lemma 2.2 *If $\{\phi_n\}$ is a sequence of functions on \mathfrak{R}^2 which converge uniformly to a bounded continuous ϕ , then*

$$\int \phi_n(l, r) dQ_n(l, r) \rightarrow \int \phi(l, r) dQ(l, r)$$

Proof of Lemma 2.2. The result follows from the following set of inequalities:

$$\begin{aligned} & \left| \int \phi_n(l, r) dQ_n(l, r) - \int \phi(l, r) dQ(l, r) \right| \\ & \leq \left| \int (\phi_n(l, r) - \phi(l, r)) dQ_n(l, r) \right| + \left| \int \phi(l, r) dQ_n(l, r) - \int \phi(l, r) dQ(l, r) \right| \\ & \leq \|\phi_n - \phi\| \int d|Q_n| + \left| \int \phi(l, r) dQ_n(l, r) - \int \phi(l, r) dQ(l, r) \right|, \end{aligned} \quad (2.4)$$

where $\|\cdot\|$ represents the supremum norm. The first term in (2.4) converges to zero since ϕ_n converges uniformly to ϕ and $\int d|Q_n| = 1$. The second term converges to zero on account of the weak convergence of Q_n to Q and ϕ being a bounded continuous function.

Proof of Theorem 2.1. The proof follows immediately by noting the following facts. P_n and Q_n converge a.s to P and Q respectively (Glivenko-Cantelli Lemma). The third term in (2.2) converges to the corresponding term in (2.1) by an application of Lemma 2.2 and the a.s. convergence of P_n and Q_n to P and Q respectively. This completes the proof of Theorem 2.1.

Theorem 2.3 (Weak Convergence) *As $n \rightarrow \infty$, the random process $Y_n(t) := \sqrt{n}(F_n^1(t) - F^1(t))$ converges weakly to a Gaussian process Y which satisfies*

$$\begin{aligned} Y(t) &= W(t) + Z(t, t) + \int_{l < t < r} \frac{P(t) - P(l)}{P(r) - P(l)} dZ(l, r) + \\ &+ \int_{l < t < r} \left[\frac{P(t)[W(l) - W(r)] + P(r)[W(t) - W(l)] + P(l)[W(r) - W(t)]}{(P(r) - P(l))^2} \right] dQ(l, r), \end{aligned} \quad (2.5)$$

where $W(t)$ is a mean zero Gaussian process with covariance function specified by $E\{W(s)W(t)\} = P(s)(A - P(t))$ for $0 \leq s < t$, where $A = P(+\infty)$, and Z is a mean zero Gaussian process with covariance $E\{Z(s, t)Z(l, r)\} = Q(s, t)(B - Q(l, r))$ for $0 \leq s < t$ and $0 \leq l < r$, where $B = Q(\infty, \infty) = P[X \in (L, R)]$, and the integral is understood to be in the Ito sense.

Proof of Theorem 2.3. Note that

$$\begin{aligned} \sqrt{n}(F_n^1(t) - F^1(t)) &= \sqrt{n}(P_n(t) - P(t)) + \sqrt{n}(Q_n(t, t) - Q(t, t)) \\ &+ \sqrt{n} \left(\int \frac{P_n(t) - P_n(l)}{P_n(r) - P_n(l)} dQ_n(l, r) - \int \frac{P(t) - P(l)}{P(r) - P(l)} dQ(l, r) \right) \end{aligned} \quad (2.6)$$

In the above equation and in what follows, the double integrals are over the region $0 < l < t < r$. Recall (see Theorem 14.3, Billingsley, 1999) that the first term above $\sqrt{n}(P_n(t) - P(t))$ converges weakly to the process W . A trivial extension of the above result to the multivariate case will imply that $\sqrt{n}(Q_n(l, r) - Q(l, r))$ will converge weakly to the process Z . Note that P_n and Q_n are assumed to be independent. We now analyze the last term in (2.6):

$$\begin{aligned} &\sqrt{n} \left(\int \frac{P_n(t) - P_n(l)}{P_n(r) - P_n(l)} dQ_n(l, r) - \int \frac{P(t) - P(l)}{P(r) - P(l)} dQ(l, r) \right) \\ &= \sqrt{n} \int \left(\frac{P_n(t) - P_n(l)}{P_n(r) - P_n(l)} - \frac{P(t) - P(l)}{P(r) - P(l)} \right) dQ(l, r) + \\ &\quad \int \frac{P_n(t) - P_n(l)}{P_n(r) - P_n(l)} d\sqrt{n}(Q_n(l, r) - Q(l, r)) \end{aligned} \quad (2.7)$$

The integrand in the first term can be written as

$$\sqrt{n} \left[\frac{[P_n(t)P(r) - P(t)P_n(r)] + [P(t)P_n(l) - P_n(t)P(l)] + [P(l)P_n(r) - P_n(l)P(r)]}{(P_n(r) - P_n(l))(P(r) - P(l))} \right] \quad (2.8)$$

The denominator converges almost surely to $(P(r) - P(l))^2$ uniformly in r and l . The first term in the numerator converges to $W(t)P(r) - W(r)P(t)$. In fact

$$\begin{aligned} \sqrt{n}[P_n(t)P(r) - P(t)P_n(r)] &= \sqrt{n}[(P_n(t) - P(t))P(r) - (P_n(r) - P(r))P(t)] \\ &\rightarrow W(t)P(r) - W(r)P(t) \end{aligned}$$

The last expression following from the fact that since $\sqrt{n}(P_n(t) - P(t))$ converges to the process W in the Skorohod space, the finite-dimensional distributions also

converge. The other two terms in the numerator in (2.8) can be analyzed similarly. Putting together and simplifying, the first term in (2.7) is seen to converge to

$$\int \frac{P(t)[W(l) - W(r)] + P(r)[W(t) - W(l)] + P(l)[W(r) - W(t)]}{(P(r) - P(l))^2} dQ(l, r) \quad (2.9)$$

Note that in the above we have used the fact that the integrand converges in the Skorohod space and the integral is a continuous functional. This is on account of the fact that the integrand in (2.9) is a process with continuous paths and this implies that the convergence of the integrand in the first term of (2.7) to it is in fact uniform (see Billingsley, 1999, pp 150).

Using integration by parts the second term on the right in (2.7), can be rewritten as

$$\int \sqrt{n}(Q_n(l, r) - Q(l, r)) d \left(\frac{P_n(t) - P_n(l)}{P_n(r) - P_n(l)} \right)$$

Using the weak convergence of the process $\sqrt{n}(Q_n - Q)$ to the continuous Gaussian process Z , and the Lemma in Billingsley (1999, pp 151), we conclude that the second term on the right in (2.7) converges to

$$\int Z(l, r) d \left(\frac{P(t) - P(l)}{P(r) - P(l)} \right) = \int \frac{P(t) - P(l)}{P(r) - P(l)} dZ(l, r),$$

where the last equality follows by integration by parts formula.

This completes the proof of Theorem 2.3.

Since the estimator F_n^1 converges to F^1 and not to the actual lifetime distribution F_0 , it is important to know how far is F^1 from the actual F_0 . The following theorem addresses this question.

Theorem 2.4 (How close is F^1 to F_0 ?)

$$(a) \quad \| F_0(t) - F^1(t) \| \leq \sup_{0 < t < \infty} [Q(t, \infty) - Q(t, t)] \quad (2.10)$$

$$(b) \quad \| F_0(t) - F^1(t) \| \leq \sup_{0 < t < \infty} \int_{l < t < r} \max(g(l, t, r), h(l, t, r)) d|Q|(l, r), \quad (2.11)$$

where

$$g(l, t, r) = \frac{Q(t, \infty) - Q(l, l)}{P(r) - P(l)} \quad (2.12)$$

and

$$h(l, t, r) = \frac{(P(t) - P(l))(Q(r, \infty) - Q(l, l))}{(P(r) - P(l))^2} \quad (2.13)$$

Remark. The bound in (a) above is easy to calculate while the one in (b) is tighter. Both the bounds can be estimated from the data by replacing P and Q by their empirical counterparts. To evaluate the supremum, we need to calculate the quantities appearing in the bounds only at the data points.

Proof of Theorem 2.4. Let

$$I(t) = \int_{u < t < v} \frac{F(t) - F(u)}{F(v) - F(u)} dQ(u, v),$$

and note that

$$I(t) \leq \int_{u < t < v} d|Q|(u, v) = Q(t, \infty) - Q(t, t). \quad (2.14)$$

From (1.4), (2.1) we get

$$|F(t) - F^1(t)| \leq \int_{l < t < r} \left| \frac{F(t) - F(l)}{F(r) - F(l)} - \frac{P(t) - P(l)}{P(r) - P(l)} \right| d|Q|(l, r) \quad (2.15)$$

The integrand in (2.15) being a difference of two conditional probabilities is bounded by 1. This together with (2.14) gives (2.10).

From (1.3), (1.4), for any $a < b$,

$$\begin{aligned} [F(b) - F(a)] - [P(b) - P(a)] &= [Q(b, b) + I(b)] - [Q(a, a) + I(a)] \\ &= \int \frac{[F(b \wedge r) - F(b \wedge l)] - [F(a \wedge r) - F(a \wedge l)]}{F(r) - F(l)} dQ(l, r) \geq 0, \end{aligned} \quad (2.16)$$

since the integrand on the right is always non-negative.

From (1.4) and (2.16), for any $l < t < r$,

$$\frac{F(t) - F(l)}{F(r) - F(l)} = \frac{P(t) - P(l) + Q(t, t) + I(t) - Q(l, l) - I(l)}{F(r) - F(l)} \quad (2.17)$$

$$\leq \frac{P(t) - P(l)}{P(r) - P(l)} + \frac{Q(t, t) + I(t) - Q(l, l) - I(l)}{P(r) - P(l)} \quad (2.18)$$

Therefore,

$$\frac{F(t) - F(l)}{F(r) - F(l)} - \frac{P(t) - P(l)}{P(r) - P(l)} \leq \frac{[Q(t, t) + I(t)] - Q(l, l)}{P(r) - P(l)} \quad (2.19)$$

Using (2.14), we get

$$\frac{F(t) - F(l)}{F(r) - F(l)} - \frac{P(t) - P(l)}{P(r) - P(l)} \leq \frac{Q(t, \infty) - Q(l, l)}{P(r) - P(l)} \quad (2.20)$$

To get the inequality in the other direction, observe that

$$\begin{aligned}
\frac{P(t) - P(l)}{P(r) - P(l)} &= \frac{(F(t) - F(l)) - [Q(t, t) - Q(l, l) + I(t) - I(l)]}{(F(r) - F(l)) - [Q(r, r) - Q(l, l) + I(r) - I(l)]} \\
&\leq \frac{F(t) - F(l)}{(F(r) - F(l)) - [Q(r, r) - Q(l, l) + I(r) - I(l)]} \\
&\leq \frac{\frac{F(t) - F(l)}{F(r) - F(l)}}{1 - \frac{Q(r, r) - Q(l, l) + I(r) - I(l)}{F(r) - F(l)}}, \tag{2.21}
\end{aligned}$$

where to get the first inequality above, we use (2.16). Simplifying the above equation, we get

$$\begin{aligned}
\frac{P(t) - P(l)}{P(r) - P(l)} - \frac{F(t) - F(l)}{F(r) - F(l)} &\leq \left[\frac{P(t) - P(l)}{P(r) - P(l)} \right] \left[\frac{Q(r, r) - Q(l, l) + I(r) - I(l)}{F(r) - F(l)} \right] \\
&\leq \frac{(P(t) - P(l))(Q(r, \infty) - Q(l, l))}{(P(r) - P(l))^2} \tag{2.22}
\end{aligned}$$

This completes the proof of Theorem 2.4.

$$F_0(t) - F^1(t) = \int_{l < t < r} \left[\frac{F_0(t) - F_0(l)}{F_0(r) - F_0(l)} - \frac{P(t) - P(l)}{P(r) - P(l)} \right] dQ(l, r) \tag{2.23}$$

For any given value of B , it can be seen from (2.23) that the discrepancy between F_0 and F^1 will be more over intervals that are heavily censored and also contain a lot of uncensored observations, and it grows with increasing B . Further, the difference between the two will persist for a longer duration as the censoring intervals become wider (see remark below). Thus, the approximation we propose will be best in situations where the censoring is light (overall and over areas which contain a lot of uncensored lifetimes) and when the length of the censoring intervals are small relative to the range of most of the lifetimes.

We consider the two examples from JM and compare their SCE with our estimator.

Example 1. Here $n = 5$ and the observations are $z_1 = 2$, $z_2 = 4$, $z_3 = 6$, $z_4 = (1, 5)$ and $z_5 = (3, 7)$. The non parametric MLE (NPML) puts mass $(5 - \sqrt{5})/10$ each on z_1 and z_3 and a mass of $1/\sqrt{5}$ on z_2 .

To derive the estimator F_n^1 we first put mass $1/5$ on each of the uncensored observations. The censored observation $(1, 5)$ contains the observations z_1 and z_2 , so the each gets an additional mass of $1/10$. The interval $(3, 7)$ contains z_2 and z_3 and so we give a mass of $1/10$ to each one of them. Thus, F_n^1 assigns a mass of $3/10$ to z_1 and z_3 , and the remaining mass of $2/5$ to z_2 . The supremum norm distance between the two estimators is $(\sqrt{5} - 2)/10$.

Example 2. Let $n = 4$ and the data be $1, 2, (3, 6)$ and $(4, 7)$. A SCE for this data is given as having mass .25 on $1, 2, 4.5, 5.5$. The NPMLE is given as having mass .25 on 1 and 2 , and a mass of .5 on some point in the overlap region $(4, 6)$ of the two censored intervals.

The approximate SCE will have a mass of .25 on $1, 2, 6, 7$.

Remark 1. Suppose that the censoring distribution Q is discrete, i.e. it puts all its mass on a countable number of intervals (l_i, r_i) , $i = 1, 2, \dots$, and these intervals are disjoint. Then, it can be easily seen that

$$\| F - F^1 \| \leq \max_i [(P(r_i) - P(l_i)) + (Q(r_i, r_i) - Q(l_i, l_i))] = \max_i [F(r_i) - F(l_i)] \quad (2.24)$$

In fact $F(x) = F^1(x)$ for all $x = l_i, r_i$, $i = 1, \dots$. As F and F^1 are non-decreasing, the result follows.

Remark 2. The approximation we propose and the results can be extended to more complex situations like spatially censored data with random censoring sets. This is being investigated and will be the content of a future work.

3 Computational Results.

A simulation study was performed to compare the performance of the ASCE with the SCE. Lifetime distributions were taken to be exponential, Erlang and Weibull. Censoring was affected by random intervals with left end points and interval width being independent exponential random variables. The results are summarized in Tables 1-3. Let F_0, E, F_n, F_n^1 denote the true lifetime distribution, the empirical distribution, the SCE and the ASCE respectively. The first column gives the parameters for generating the censoring intervals and the second for the lifetimes. The third column gives the percentage of observations that were censored. Sample size in all experiments was taken to be 100. The fourth column compares the ASCE with the SCE, while the subsequent ones compare the empirical and the ASCE with the actual distribution F_0 respectively, comparisons being in terms of the sup norm distance. Lifetime distributions are taken to be exponential, Gamma and Weibull respectively in Tables 1-3. It can be seen that the ASCE is very close to the SCE. 20 iterations were used to compute the SCE. The censoring percentage is seen to vary between 16 and over 50 percent. Still the ASCE is seen to perform well in comparison with the EDF (which in practice would be unknown).

Cen. Par.	Lif. Par.	% Cen.	$\ F_n - F_n^1 \ $	$\ F_0 - E \ $	$\ F_0 - F_n^1 \ $
0.2,0.2	0.1	20	0.0045	0.1072	0.1594
0.02,0.02	0.01	18	0.0027	0.0727	0.0630
0.02,0.02	0.05	16	0.0067	0.1252	0.1539
0.05,0.05	0.05	24	0.0043	0.064	0.0820
0.1,0.01	0.05	51	0.0604	0.0682	0.1352

Table 1.

Cen. Par.	Lif. Par.	% Cen.	$\ F_n - F_n^1 \ $	$\ F_0 - E \ $	$\ F_0 - F_n^1 \ $
0.1,0.1	2,1	13	0.0026	0.0562	0.0573
0.1,0.1	5,1	21	0.0038	0.0485	0.0596
0.1,0.01	5,1	36	0.0308	0.0821	0.1420

Table 2.

Cen. Par.	Lif. Par.	% Cen.	$\ F_n - F_n^1 \ $	$\ F_0 - E \ $	$\ F_0 - F_n^1 \ $
0.1,0.1	5,2	34	0.0063	0.0931	0.1312
0.1,0.01	10,2	57	0.1349	0.0544	0.1051

Table 3.

Finally, we consider an actual data set on melanoma survival collected at Odense University Hospital, Denmark (see Anderson *et. al* (1993)). The sample contains 205 data points, ranging from 10 to 5565. The data were censored by random intervals of the form $(L, L + W)$ where L and W are independent exponential random variables with means 2000 and 1000 respectively. This resulted in roughly 18% of the observations being censored. The supremum norm distance between the EDF and the ASCE was only 0.0199. This is quite remarkable given that 18% of the data is censored.

4 Conclusion

We propose and study the performance of a simple alternative estimator for lifetime distribution for middle-censored data. The main advantages of this simpler estimator are that it is not recursive, and consistency and weak convergence can be established. We verify through simulations that it performs in practice as well as the SCE proposed in Jammalamadaka and Mangalam (2003).

References

- [1] Anderson, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993), *Statistical models based on counting processes*, Springer-Verlag, New York.
- [2] Billingsley, P., (1999), *Convergence of probability measures, 2nd Ed.*, Wiley, New York.
- [3] Chang, M.N., (1990), Weak Convergence of a self-consistent estimator of the survival function with doubly censored data, *Annals of Statistics* **18**, 391-404.
J. Roy. Statist. Soc. Ser. B, **39**, 1-38.
- [4] Gehan, Ehmund A. (1965), A generalized two-sample Wilcoxon test for doubly censored data, *Biometrika*, **52**, 650-653.
- [5] Geskus, R.B. and Groeneboom, P. (1996), Asymptotically optimal estimation of smooth functionals for interval censoring, *I. Statist. Neerlandica*, **50**, 69-88.
- [6] Groeneboom, P. and Wellner, J.A. (1992), Information bounds and non-parametric maximum likelihood estimation, *DMV seminar*, **19**, Birkhauser Verlag, Basel.
- [7] Jammalamadaka, S.Rao, and Mangalam, V. (2003), Nonparametric estimation for middle censored data, to appear in *Jour. of Nonparametric Statistics*, **15**, No.3.
- [8] Kaplan, E.L. and Meir, P. (1958), Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**, 457-481.
- [9] Tarpey, T., and Flury, B. (1996), Self-consistency: A fundamental concept in statistics. *Stat. Science*, **11**, 229-243.
- [10] Tsai, W.Y., and Crowley, J. (1985), A large sample study of generalised maximum likelihood estimators from incomplete data via self-consistency, *Ann. Stat.*, **13**, 1317-1334.
- [11] Turnbull, B.W. (1974), Nonparametric estimation of survivor function with doubly censored data, *J. Amer. Stat. Assoc.*, **69**, 169-173.