東海大學統計研究所

碩士論文

A Self-Consistent Estimator for Interval-Censored and Left-Truncated Data

指導教授：沈葆聖博士

研究生：陳咨瑋

中華民國一〇〇年七月

A Self-Consistent Estimator for Interval-Censored and Left-Truncated Data

Director :Pao-sheng Shen

Student: Tuz-Wei Chen

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

致謝

經過一年的研究，終於完成了碩士論文。一年前一心只想著做一個從事統計相關研究的研究生，因此沈葆聖教授成了我的第一首選，在許多次的學習中不斷的磨練自己的統計能力，在許多的課業中增進自己的知識，更在許多事情中學習到處理時情的方法及態度。雖然這一年中經歷了模擬的低潮，許多程式結果不理想，甚至對自己沒信心，但也都在一次次指導教授及大家的鼓勵下，一點一滴的完成到目前就究成果。

首先，我要感謝沈教授這一年的教導與照顧，讓老師在我的論文上操了多心，幫助我在學習上能有更好的效率。另外我要感謝研究室中教導過我許多細節的學長姐們。並且感謝怡穎和雨青在一路上的互相打氣並幫我分擔了許多生活的細節，感謝炳然、金強、志維在我學習遇到瓶頸時，熱心的幫助及鼓勵我，及許多系上一直幫我加油的同學們和學弟妹，有妳們的鼓勵和歡笑讓我心情能保持愉快。

此外，我還要感謝我的家人，體諒我因為研究生活沒有好好照顧家裡的情形，還有父親的體貼並不時的給予我鼓勵、建議和提醒我讓我能完成這份碩士學位。

這本論文將獻給每一位幫助過我的家人、朋友和研究室的夥伴們，我會以此為動力，更努力的在未來有所表現。

Abstract

Interval censoring refers to a situation in which, T, the time to occurrence of an event of interest is only known to lie in an interval $[L, R]$. In some cases, the variable T also suffers left-truncation. Based on an integral equation, we propose a self-consistent estimator (SCE) of survival function of T. It is shown that the NPMLE is a solution of the integral equation. Under some conditions, we show the consistency of the SCE.

Key Words: left truncation; interval censoring; self-consistent.

Content

1. Introduction

Left truncated and interval-censored data often arise in epidemiology and individual follow-up studies and possibly in other fields. Their importance stems from the common use of prevalent cohort study designs to estimate survival from onset of a specified disease. Consider the following example.

Example 1: AIDS Cohort Studies

In AIDS cohort studies, we are interested in the incubation time of the disease. An individual is selected only when he (or she) is HIV-positive and yet none have developed AIDS. Hence, earlier onset of AIDS would then be a truncating force for the variable of interest. Suppose that the infection time (denoted by T_s) can be quite accurately determined (e.g. due to blood transfusion). The recruitment starts at τ_0 and the follow-up is terminated at τ_e. For each individual i, let T_i^* denote the time from T_s to development of AIDS. Let $V_i^* = \tau_0 - T_s$ if $T_s < \tau_0$ and $V_i^* = 0$ if $T_s \geq \tau_0$. Let $C_i^* = \tau_e - T_s$ denote the censoring times. Furthermore, there are many situations, in which the onset of AIDS is recorded only between an interval although the initiating events (HIV infection) T_s is recorded exactly. Hence, the variable of interest T_i^* is only recorded between an interval, say $[L_i^*, R_i^*]$. Note that when T_i^* is right censoring, we can write $[L_i^*, R_i^*]$ as $[C_i^*, \infty]$. In this case, T_i^* is subject to left-truncated and interval-censored. Hence, one observes nothing if $T_i^* < V_i^*$, and observes $([L_i^*, R_i^*], V_i^*)$ if $T_i^* \geq V_i^*$. We assume that T_i^* is independent of (V_i^*, L_i^*, R_i^*) and V_i^* is dependent of (L_i^*, R_i^*) with $P(V_i^* \leq L_i^* | T_i^* \geq V_i^*) = 1$.

Let $F(t)$ denote the distribution function of T_i^*, and $G(x)$and $Q(x)$ denote the distribution function of V_i^* and C_i^*, respectively. For any distribution function W denote the left and right endpoints of its support by $a_W = inf\{t : W(t) > 0\}$ and $b_W = inf\{t : W(t) = 1\}$, respectively. Throughout this article we assume that T_i^*, L_i^*, R_i^* and V_i^* are all continuous, and

$$a_G \leq a_F \quad \text{and} \quad b_G \leq b_F \leq b_Q. \tag{1.1}$$

Let $(L_1, R_1, V_1), \ldots, (L_n, R_n, V_n)$ denote the left-truncated and interval-censored data. Note that $[L_i, R_i] \subset [V_i, \infty]$, i.e. $V_i \leq L_i$. The nonparametric maximum likelihood estimator (NPMLE) of F can be obtained by using EM algorithm of Turnbull (1976). When there is no truncation, the asymptotic properties of the NPMLE have been derived for interval-censored data. Groeneboom and Wellner (1992) proposed an iterative convex minorant algorithm to calculate the NPMLE and proved the uniform consistency of the NPMLE when F is continuous and the joint distribution function of (L, R) is absolutely continuous. If (L, R) is assumed discrete, the NPMLE has the usual $\sqrt{n}$ convergence rate and a normal limiting distribution (Yu et al. (1998a, b)). However, if (L, R) is continuous, the NPMLE converges slower than $\sqrt{n}$ to a non-Gaussian limiting distribution (see Groeneboom and Wellner (1992), Shick and Yu (2000), van der Vaart and Wellner (2000), Song (2004)). Although asymptotic properties of the NPMLE have been derived for the interval-censored data without truncation much less is known about the large sample properties of the NPMLE if both interval censoring and truncation are present. Pan and Chappell (1999) showed that the NPMLE is inconsistent when data is subject to case 1 interval censoring and left truncation. Under the assumption of monotonic hazard function, Pan et al. (1998) showed the consistency of the NPMLE when data is subject to left truncation and interval censoring.

In Section 2, based on an integral equation, we propose a self-consistent estimator (SCE) of survival function of T_i^*. We show that the NPMLE is a solution of the proposed integral equation. Under some conditions, we show the consistency of the SCE. In Section 3, a simulation study is conducted to compare the performance between the SCE and NPMLE.

2. The Nonparametric Estimators

2.1 The NPMLE

In this section, we briefly review the NPMLE of $S_F(t) = P(T_i^* > t)$ using EM algorithm of Turnbull (1976). Notice that due to sampling scheme described in Section 1, we have $P([L_i, R_i] \subset [V_i, \infty)) = 1$. Without loss of generality, suppose the observed data are ordered according to L_i such that $L_1 < L_2 < \cdots < L_n$. Following Turnbull (1976), Frydman (1994) and Alioum and Commenges (1996), we consider nonparametric estimation of F using the n independent pairs $\{A_1, B_1\}, \ldots, \{A_n, B_n\}$, where $A_i = [L_i, R_i]$ and $B_i = [V_i, \infty)$. Assuming that the inspection process which gives rise to A_i is independent of T_i, we consider the following conditional likelihood:

$$L_c(S_F) = \prod_{i=1}^{n} \frac{P_{S_F}(A_i)}{P_{S_F}(B_i)}, \tag{2.1}$$

where $P_S(R)$ denotes the probability that is assigned to the interval by S_F. We define an NPMLE as $\hat{S}_M = \operatorname{argmax}_{S \in \mathcal{S}}\{L_c(S)\}$, where $\mathcal{S}$ denotes the class of survival functions such that $P_S(\cup_{i=1}^n B_i) = 1$ and $L_c(S)$ is defined, i.e. $P_S(B_i) > 0$ for all $i = 1, \ldots, n$. Using the approach of Hudgens (2005), we define $\mathcal{K} = \{K_1, K_2, \ldots, K_{2n}\}$, where $K_1 = A_i$ for $i = 1, \ldots, n$, and $K_i = (-\infty, V_i)$ for $i = n+1, \ldots, 2n$. An intersection graph for $\mathcal{K}$ is constructed as follows. For each element of $\mathcal{K}$, we define a corresponding vertex. Let i be the label of the vertex corresponding to K_i. Denote the set of vertex by S_v. Two vertices in S_v are considered connected by an edge if and only if the two corresponding regions in $\mathcal{K}$ intersect. A clique is defined as a subset M of S_v such that every member of M is connected by an edge to every other member of M. A maximal clique has the additional property that it is not a proper subset of any other clique. Let $\mathcal{M} = \{M_1, \ldots, M_J\}$ be the subset of maximal cliques of S_v that contain at least one vertex corresponding to a censoring interval, i.e. for each $M_j \in \mathcal{M}$, there is some $i \in \{1, \ldots, n\}$ such that $i \in M_j$. Let $\mathcal{H} = \{H_1, \ldots, H_J\}$ be the corresponding set of real representations of elements of $\mathcal{M}$ where $H_j = \cap_{i \in M_j} K_i$ for $j = 1, \ldots, J$. For example, when $A_1 = [2, 5], B_1 = [1, \infty)$, $A_2 = [4, 9]$, $B_2 = [1, \infty)$, $A_3 = [6, 7]$, and $B_3 = [3, \infty)$, we obtain $H_1 = [2, 3]$, $H_2 = [4, 5]$, and $H_3 = [6, 7]$. By Lemma

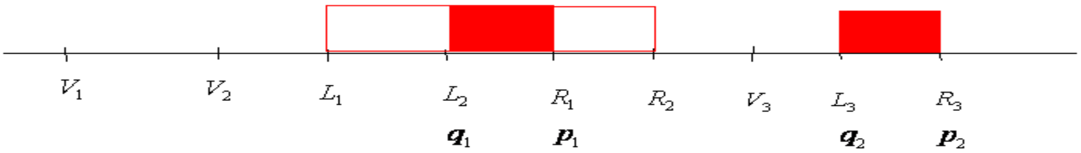

Figure 1. Schematic depiction of the innermost set $[q_j, p_j]$

1 of Hudgens (2005), any distribution function which increases outside $\cup_{j=1}^{J} H_j$ cannot be an NPMLE. By Lemma 2 of Hudgens (2005), for fixed value of $P_F(H_j)$, the likelihood is independent of the values of F within the region H_j. These lemmas allow us to consider maximizing a simpler likelihood than equation (2.1). For each $H_j \in \mathcal{H}$, let $s_j = P_F(H_j)$ and let $\mathbf{s}$ be an m-dimension column vector with elements s_j. We shall assume throughout that $H_1, \ldots, H_J$ are ordered such that $H_j = [q_j, p_j]$ is to the left of $H_{j+1} = [q_{j+1}, p_{j+1}]$ for $j = 1, \ldots, J-1$, i.e. $[q_1, p_1]$, $[q_2, p_2], \ldots, [q_J, p_J]$, where $q_1 \le p_1 < q_2 \le p_2 < \cdots < q_J \le p_J$. Figure 1 also highlights the innermost set $[q_j, p_j]$. It follows that from lemmas 1 and 2 of Hudgens (2005) that maximizing likelihood (2.1) is equivalent to maximizing

$$L_c(\mathbf{s}) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{J} \alpha_{ij} s_j}{\sum_{j=1}^{J} \beta_{ij} s_j}, \tag{2.2}$$

where $\alpha_{ij} = I[H_j \subset A_i]$, $\beta_{ij} = I[H_j \subset B_i]$ and $I[\cdot]$ is the usual indicator function. The resulting reduced likelihood (2.2) is exactly as described in section 2 of Alioum and Commenges (1996). The goal is to maximize likelihood (2.2) subject to the constraints

$$\sum_{j=1}^{J} s_j = 1, \tag{2.3}$$

$$s_j \ge 0 \ (j = 1, \ldots, J), \tag{2.4}$$

and

$$\sum_{j=1}^{J} \alpha_{ij} s_j > 0,\ \ (i = 1, \dots, n). \tag{2.5}$$

We shall use Ω to denote the parameter space that is given by constraints (2.3)-(2.5), i.e.

$$\Omega = \{\mathbf{s} \in R^J : \sum_{j=1}^{J} s_j = 1; s_j \geq 0 \text{ for } j = 1, \dots, J; \sum_{j=1}^{J} \alpha_{ij} s_j > 0 \text{ for } i = 1, \dots n\}.$$

To find the maximum likelihood estimate of the vector $\mathbf{s}$, we can use an EM algorithm and the resulting self-consistent estimate of $\mathbf{s}$ is exactly the Turnbull's (1976) self-consistency algorithm as follows:

$$s_j^{(b)} = \left\{1 + \frac{d_j(s^{(b-1)})}{M(s^{(b-1)})}\right\} s_j^{(b-1)}\ (1 \leq j \leq J), \tag{2.6}$$

where

$$d_j(s^{(b-1)}) = \sum_{i=1}^{n} \left\{ \left(\alpha_{ij} \Big/ \sum_{k=1}^{J} \alpha_{ik} s_k^{(b-1)} \right) - \left(\beta_{ij} \Big/ \sum_{k=1}^{J} \beta_{ik} s_k^{(b-1)} \right) \right\},$$

and

$$M(s^{(b-1)}) = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{J} \beta_{ij} s_j^{(b-1)}}.$$

Let $\hat{s}_j$ $(j = 1, \dots, J)$ denote the estimators obtained from (2.6). As pointed out by Hudgens (2005), in general, a maximizer of $L_c(\mathbf{s})$ subject to $s \in \Omega$ need not exist since Ω is not closed. For left-truncated and interval-censored data, Hudgens (2005) (see Theorem 1, page 578) proposed a sufficient and necessary condition for the existence of the NMPLE as follows:

" There is a maximizer of $L_c(\mathbf{s})$ subject to $\mathbf{s} \in \Omega$ if and only if for each non-empty proper subset $\mathcal{S}$ of $\{1, \dots, n\}$ there is an $i \notin \mathcal{S}$ such that $\mathcal{A}_i \subset \mathcal{D}_S$, $\mathcal{A}_i = \cup_{j \in A_i^*} H_j$, $\mathcal{D}_S = \cup_{k \in S} \mathcal{B}_k$, $\mathcal{B}_k = \cup_{j \in B_k^*} H_j$, where $A_i^* = \{j : \alpha_{ij} = 1\}$ and $B_k^* = \{j : \beta_{kj} = 1\}$". Based on the estimators $\hat{s}_j$'s, an estimator $\hat{S}_M(t)$ of $S_F(t)$ can be uniquely defined for $t \in [p_j, q_{j+1})$ by $\hat{S}_M(p_j) = \hat{S}_M(q_{j+1}-) = 1 - (\hat{s}_1 + \cdots + \hat{s}_j)$, but is not uniquely defined for t being in an open innermost interval (q_j, p_j) with $q_j < p_j$. To avoid ambiguity we define $\hat{S}_M(t) = 1 - [\hat{s}_1 + \cdots + \hat{s}_{j-1} + s_j(t - q_j)/(p_j - q_j)]$ if $t \in (q_j, p_j]$ and $0 < q_j < p_j < \infty$. Figure 2 highlights the estimated distribution function $F_M(t) = 1 - S_M(t)$.

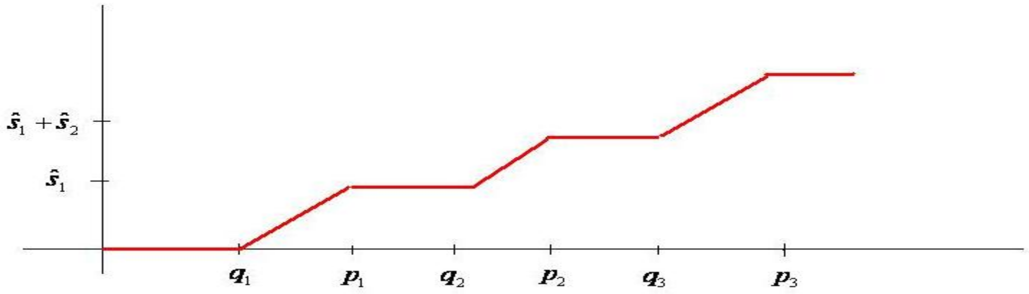

Figure 2. Schematic depiction of the estimated d.f. F_M

2.2 The SCE

Let $S_F(t) = 1 - F(t)$ denote the survival function of T and $p = P(V_i^* \leq T_i^*)$ denote the proportion of un-truncation. We have the following equation:

$$S_F(t) = P(T_i^* > t, V_i^* \leq t) + P(T_i^* > t, V_i^* > t)$$

$$= pP(V_i^* \leq t < L_i^* | T_i^* \geq V_i^*) + pP(T_i^* > t, L_i^* < t \leq R_i^* | T_i^* \geq V_i^*) + P(T_i^* > t, V_i^* > t). \tag{2.7}$$

Motivated by (2.7), given p, we consider the following self-consistent estimator:

$$\hat{S}(t) = \frac{1}{np^{-1}} \left\{ \sum_{i=1}^{n} I_{[V_i \leq t < L_i]} + \sum_{i=1}^{n} I_{[L_i \leq t < R_i]} \frac{\hat{S}(t) - \hat{S}(R_i)}{\hat{S}(L_i) - \hat{S}(R_i)} + \sum_{i=1}^{n} I_{[V_i > t]} \frac{\hat{S}(t)}{\hat{S}(V_i)} \right\}. \tag{2.8}$$

Notice that the last term of the equation (2.8) is to recover the missing information due to left-truncation. Given the observation $V_i > t$, a pseudo observation is recovered by adding the weight $\hat{S}(t)/\hat{S}(V_i)$. Let $\tilde{G}(t) = P(V_i \leq t)$ denote the sub-distribution function of V_i. Since $\tilde{G}(t) = p^{-1} \int_0^t 1/S_F(V_i) dG(t)$. It follows that np^{-1} can be estimated by $\sum_{i=1}^n 1/S_F(V_i)$ (see Shen (2005)). Hence, a self-consistent estimator $\hat{S}_n$ is given by solving the following equation:

$$\hat{S}_n(t) =$$

$$\left[\sum_{i=1}^{n} \frac{1}{\hat{S}_n(V_i)} \right]^{-1} \left\{ \sum_{i=1}^{n} I_{[V_i \leq t < L_i]} + \sum_{i=1}^{n} I_{[L_i \leq t < R_i]} \frac{\hat{S}_n(t) - \hat{S}_n(R_i)}{\hat{S}_n(L_i) - \hat{S}_n(R_i)} + \sum_{i=1}^{n} I_{[V_i > t]} \frac{\hat{S}_n(t)}{\hat{S}_n(V_i)} \right\}. \tag{2.9}$$

Let $\tilde{G}_n(v)$ denote the empirical version of $\tilde{G}(v)$. Similarly, Let $\tilde{H}_n(v,l)$ and $\tilde{Q}_n(l,r)$ denote the empirical versions of the joint sub-distributions of $\tilde{H}(v,l) = P(V_i \le v, L_i \le l)$ and $\tilde{Q}(l,r) = P(L_i \le l, R_i \le r)$, respectively. It follows that (2.9) can be written as

$$\hat{S}_n(t) =$$

$$\left[\int \frac{1}{\hat{S}_n(v)}\tilde{G}_n(dv)\right]^{-1}\left\{\int_{v\le t<l}\tilde{H}_n(dv,dl)+\int_{l\le t<r}\frac{\hat{S}_n(t)-\hat{S}_n(r)}{\hat{S}_n(l-)-\hat{S}_n(r)}\tilde{Q}_n(dl,dr)+\int_{v>t}\frac{\hat{S}_n(t)}{\hat{S}_n(v)}\tilde{G}_n(dv)\right\}. \tag{2.10}$$

Notice that when there is no truncation, (2.10) is reduced to the following self-consistent equation:

$$\hat{S}_n(t) = \int_{t<l}\tilde{Q}_{L,n}(dl) + \int_{l\le t<r}\frac{\hat{S}_n(t)-\hat{S}_n(r)}{\hat{S}_n(l)-\hat{S}_n(r)}\tilde{Q}_n(dl,dr), \tag{2.11}$$

where $\tilde{Q}_{L,n}$ is the empirical version of the sub-distribution function of $\tilde{Q}_L(l) = P(L_i \le l)$. Note that equation (2.11) is the same as equation (2.2) of Yu et al. (2001) for mixed interval censored data.

The following theorem shows that $\hat{S}_M$ satisfies the equation (2.9).

Theorem 1.

The NPMLE $\hat{S}_M$ satisfies equation (2.9).

Proof:

First, consider an initial estimator $\hat{S}_n^{(0)}$, which puts mass only on $[q_j, p_j]$ $(j = 1,\ldots,J)$. Let $\hat{S}_n^{(1)}$ denote the first step estimator. Without changing the innermost intervals and likelihood function, we can transform data by moving all right censored and left truncated points between p_{j-1} and q_j to p_{j-1}. Similarly, move all left censored points between p_{j-1} and q_j to q_j. (see Li et al. (1997)). Based on the transform data, for all i,j, we have $I_{[p_{j-1}<V_i\le q_j]} = 0$, $I_{[V_i\le p_{j-1}\le L_i]}I_{[q_j>L_i]} = 0$, $I_{[V_i\le p_{j-1}\le L_i]}I_{[q_j>L_i]} = 0$, $I_{[V_i>p_{j-1}]}I_{[V_i\le q_j-\le L_i]} = 0$, $I_{[L_i\le p_{j-1}<R_i]} = 0$ and $I_{[L_i\le q_j-\le R_i]} = 0$. It follows that $\hat{S}_n^{(1)}(p_{j-1}) - \hat{S}_n^{(1)}(q_j-) = 0$. Hence, $\hat{S}_n^{(1)}$ also puts mass only on $[q_j,p_j]$ $(j = 1,\ldots,J)$. Next, since there is no left censoring observations in $(q_j,p_j]$ and there is no left truncation observations in $[q_j,p_j)$, we have for all i,j, $I_{[V_i\le q_j<L_i]}I_{[p_j\ge L_i]} = 0$ and $I_{[V_i>q_j]}I_{[V_i\le p_j<L_i]} = 0$. Furthermore, given an interval $[L_i,R_i]$, we either have $[q_j,p_j] \subseteq [L_i,R_i]$ or $[q_j,p_j] \cap [L_i,R_i] = \emptyset$. Thus, we have

$$\hat{S}_n^{(1)}(q_j-) - \hat{S}_n^{(1)}(p_j) = \left[\sum_{i=1}^{n}\frac{1}{\hat{S}_n^{(0)}(V_i)}\right]^{-1}\left\{\sum_{i=1}^{n} I_{[[q_j,p_j]\in([L_i,R_i]]}\frac{\hat{S}_n^{(0)}(q_j-)-\hat{S}_n^{(0)}(p_j)}{\hat{S}_n^{(0)}(L_i)-\hat{S}_n^{(0)}(R_i)}\right.$$

$$+\sum_{i=1}^{n}\frac{\hat{S}_n(q_j-)-\hat{S}_n(p_j)}{\hat{S}_n(V_i)}-\sum_{i=1}^{n}I_{[V_i\leq q_j]}\frac{\hat{S}_n(q_j-)}{\hat{S}_n(V_i)}+\sum_{i=1}^{n}I_{[V_i\leq p_j]}\frac{\hat{S}_n(p_j)}{\hat{S}_n(V_i)}\Bigg\}. \qquad (2.12)$$

Since there is no left truncation observations in $[q_j, p_j)$, (2.12) can be written as

$$\hat{S}_n^{(1)}(q_j-)-\hat{S}_n^{(1)}(p_j)=\left[\sum_{i=1}^{n}\frac{1}{\hat{S}_n^{(0)}(V_i)}\right]^{-1}\Bigg\{\sum_{i=1}^{n}I_{[[q_j,p_j]\in([L_i,R_i]]}\frac{\hat{S}_n^{(0)}(q_j-)-\hat{S}_n^{(0)}(p_j)}{\hat{S}_n^{(0)}(L_i)-\hat{S}_n^{(0)}(R_i)}$$

$$+\sum_{i=1}^{n}\frac{\hat{S}_n(q_j-)-\hat{S}_n(p_j)}{\hat{S}_n(V_i)}-\sum_{i=1}^{n}I_{[q_j\geq V_i]}\frac{\hat{S}_n(q_j-)-\hat{S}_n(p_j)}{\hat{S}_n(V_i)}\Bigg\}. \qquad (2.13)$$

Next,

$$\hat{S}_M(q_j-)-\hat{S}_M(p_j)=\left[\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{J}\beta_{ij}s_j}\right]^{-1}\Bigg\{\sum_{i=1}^{n}\frac{\alpha_{ij}}{\sum_{k=1}^{J}\alpha_{ik}\hat{s}_k}+\sum_{i=1}^{n}\frac{1-\beta_{ij}}{\sum_{k=1}^{J}\beta_{ik}\hat{s}_k}\Bigg\}\hat{s}_j. \qquad (2.14)$$

By definitions of A_i, B_i, α_{ij} and β_{ij}, it follows that equation (2.13) is equivalent to equation (2.14). The proof is completed.

The following Theorem shows the consistency of $\hat{S}_n(t)$.

Theorem 2.

Let $\Theta = \{f : f \text{ is nonincreasing function from } [a_F, b_F] \text{ to } [0,1], f(b_F) = 0 \text{ and } f(a_F) = 1\}$. Let $S \in \Theta$ be a $[0,1]$-valued function and h be a function such that $h(t)K(t) =$

$$S(t)\int_{v\leq t}\frac{h(v)}{S(v)}G(dv)-\int_{l\leq t<r}\frac{h(l)[S(t)-S(r)]}{S(l)-S(r)}Q(dl,dr)-\int_{l\leq t<r}\frac{h(r)[S(l)-S(t)]}{S(l)-S(r)}Q(dl,dr), \qquad (2.15)$$

where $h(t) = S(t) - S_F(t)$ and $K(t) = G(t) - P(L_i^* \leq t < R_i^*)$. (i) Suppose that (2.15) holds on the set $\{t : 0 < S(t) < 1\}$ implies that $h(t) = 0$ for all $t \in (a_F, b_F)$ provided that $h(t+) \neq h(t) \Rightarrow S(t+) < S(t)$ on $\{t : 0 < S(t) < 1\}$, $h(t) = 0$ on $\{t : S(t) = 0 \text{ or } S(t) = 1\}$. (ii) Suppose that $\hat{S}_n \in \Theta$ and $\hat{S}_n$ is right continuous, then $\lim_{n\to\infty}\sup_{a_F<x<b_F}|\hat{S}_n(x) - S_F(x)| = 0$ a.s.

Proof:

Let Ω be the event $\{\lim \tilde{H}_n(v,l) = \tilde{H}(v,l), \lim \tilde{Q}_n(l,r) = \tilde{Q}(l,r) \text{ uniformly for all} v < l < r\}$. For each $\omega \in \Omega$, let $\hat{S}_n$ be the solution of (2.9). Since $\{\hat{S}_n\}_{n\geq 1}$ is bounded and monotone, for each subsequence of natural numbers, by Helly's selection theorem, there exists a further subsequence, say $\{n_k\}$, such that $\lim_{n_k\to\infty}\hat{S}_{n_k}(t) = S_0(t)$ pointwisely for some $S_0 \in \Theta$. Thus, it suffices to show that $S_0(t) = S_F(t)$ for all $t \in [a_F, b_F]$.

Since $\tilde{H}_n$ and $\tilde{Q}_n$ converge uniformly to $\tilde{H}$ and $\tilde{Q}$, respectively and $\hat{S}_n$ satisfies (2.9), by the bounded convergence theorem S_0 satisfies the following equation: $S_0(t) =$

$$\left[\int \frac{1}{S_0(v)}\tilde{G}(dv)\right]^{-1}\left\{\int_{v\le t<l} d\tilde{H}(v,l) + \int_{l\le t<r} \frac{S_0(t)-S_0(r)}{S_0(l)-S_0(r)}\tilde{Q}(dl,dr) + \int_{v>t}\frac{S_0(t)}{S_0(v)}\tilde{G}(dv)\right\}. \tag{2.16}$$

Equation (2.16) can be written as

$$S_0(t)\int_{v\le t}\frac{1}{S_0(v)}\tilde{G}(dv) = \int_{v\le t<l}\tilde{H}(dv,dl) + \int_{l\le t<r}\frac{S_0(t)-S_0(r)}{S_0(l)-S_0(r)}\tilde{Q}(dl,dr). \tag{2.17}$$

Let $H(v,l) = P(V_i^* \le u, L_i^* \le l)$ and $Q(l,r) = P(L_i^* \le r, R_i^* \le r)$. Since $\tilde{G}(dv) = p^{-1}S_F(v)G(dv)$, $\tilde{H}(dv,dl) = p^{-1}S_F(l)H(dv,dl)$, and $\tilde{Q}(dl,dr) = p^{-1}[S_F(l)-S_F(r)]Q(dl,dr)$, (2.17) can be written as

$$p^{-1}S_0(t)\int_{v\le t}\frac{S_F(v)}{S_0(v)}G(dv) = p^{-1}\int_{v\le t<l} S_F(l)H(dv,dl)$$

$$+p^{-1}\int_{l\le t<r}\frac{S_0(t)-S_0(r)}{S_0(l)-S_0(r)}[S_F(l)-S_F(r)]Q(dv,dl,dr). \tag{2.18}$$

Replacing $S_0(\cdot)$ of (2.18) by $S_F(\cdot)$, we obtain

$$p^{-1}S_F(t)G(t) = p^{-1}\int_{v\le t<l} S_F(l)H(dv,dl) + p^{-1}\int_{l\le t<r}[S_F(t)-S_F(r)]Q(dl,dr). \tag{2.19}$$

Note that (2.19) is equivalent to

$$P(T_i^* > t, V_i^* \le t | T_i^* \ge V_i^*) = P(V_i^* \le t < L_i^* | T_i^* \ge V_i^*) + P(T_i^* > t, L_i^* < t < R_i^* | T_i^* \ge V_i^*).$$

Subtracting (2.19) from (2.18), we obtain

$h(t)K(t) =$

$$S_0(t)\int_{v\le t}\frac{h(v)}{S_0(v)}G(dv) - \int_{l\le t<r}\frac{h(l)[S_0(t)-S_0(r)]}{S_0(l)-S_0(r)}Q(dl,dr) - \int_{l\le t<r}\frac{h(r)[S_0(l)-S_0(t)]}{S_0(l)-S_0(r)}Q(dl,dr),$$

where $h(t) = S_0(t) - S_F(t)$. By assumption (i), it follows that $h(t) = 0$ for $t \in [a_F, b_F]$. It follows that $S_0(t) = S_F(t)$ for all $t \in [a_F, b_F]$. By (2.3) all limit points of $\hat{S}_n$ must satisfy (2.5), by Helly-Bray selection theorem we have $\hat{S}_n(t) \to S_F(t)$ a.s. for $t \in (a_F, b_F)$. Since $\hat{S}_n$ is a sequence of monotone, right continuous and bounded functions on (a_F, b_F), it follows that $\sup_{t\in(a_F,b_F)} |\hat{S}_n(t) - S_F(t)| \to 0$ a.s. (see Proposition 3.1 of Yu et al. (2001))

The proof is completed.

3. Simulation Results

A simulation study is conducted to investigate the performance of the proposed estimator $\hat{F}(t)$. The T_i^*'s are i.i.d. exponential distributed with mean equal to 1. The V_i^*'s are i.i.d. exponential distributed with scale parameters $\theta = 0.5, 1$ and 2, i.e. $G(x;\theta) = 1 - \exp(-\theta x)$ for $x > 0$. The T_i^* and V_i^* are independent to each other. To make the truncated sample interval-censored, we first generate a random variable $X = 2 + B(n_c, 0.5)$, where $B(n_c, 0.5)$ is a binomial random variable with $n_c = 4, 6$. Given $X = k$, we then generate k i.i.d uniform random variables $U_{ji} \sim U(0,1)$ $(j = 1, \ldots, k)$. Define $Z_{1i} = V_i^* + U_{1i}$, $Z_{2i} = U_{2i} + Z_{1i}$, $Z_{3i} = U_{3i} + Z_{2i}$, $\cdots$, $Z_{ki} = Z_{k-1,i} + U_{ki}$. We keep the sample if $T_i^* \geq V_i^*$ and regenerate a sample if $T_i^* < V_i^*$. If T_i^* falls in the interval $[Z_{ji}, Z_{j+1,i}]$ $(j = 1, \ldots, k-1)$, then let $L_i^* = Z_{ji}$ and $R_i^* = Z_{j+1,i}$. If $T_i^* > Z_{k,i}^*$ then let $L_i^* = Z_{k,i}$ and $R_i^* = 10000$. The goal is to estimate $S(t_p) = p$, with $p = 0.8$, 0.5 and 0.2. The sample sizes n are chosen as 200 and 400. Based on left-truncated and interval-censored data (V_i, L_i, R_i) $(i = 1, \ldots, n)$, we obtain the proposed estimator $\hat{S}_n(t_p)$ and the NPMLE $\hat{S}_M(t_p)$. The sample sizes are chosen as 200 and 400. The replication is 1000 times. Tables 1 through 3 show the empirical biases, standard deviations (std.) and mean squared errors (mse) of $\hat{S}_n$ and $\hat{S}_M$. Tables 1 through 3 also list the proportion of truncation $P(T_i^* < V_i^*)$ (denoted by q_T). Based on the results of Tables 1 through 3, we conclude that:

(i) Given q_T, the rmse of the estimators $\hat{S}_n$ and $\hat{S}_M$ increase as n_c decreases, i.e. mean interval length increases.

(ii) Given n_c, the rmse of estimators $\hat{S}_n$ and $\hat{S}_M$ increase as proportion of truncation q_T increases.

(iii) In terms of rmse, when $n = 200$ the NPMLE $\hat{S}_M$ outperforms the SCE $\hat{S}_n$. When $n = 400$, the performance of the estimators $\hat{S}_n$ and $\hat{S}_M$ are close to each other for most of cases considered.

Table 1. Simulation results for bias, standard deviation and root mean squared error for estimating $S(t_{0.2})$

				$\hat{S}_n(t_{0.2})$			$\hat{S}_M(t_{0.2})$		
θ	n_c	n	q_T	bias	std	rmse	bias	std	rmse
2	5	200	0.43	-0.074	0.023	0.072	-0.065	0.021	0.068
2	5	400	0.43	-0.050	0.017	0.060	-0.056	0.014	0.057
2	6	200	0.43	0.074	0.026	0.078	-0.077	0.023	0.080
2	6	400	0.43	0.059	0.015	0.061	-0.062	0.013	0.063
4	5	200	0.31	-0.061	0.024	0.065	-0.054	0.023	0.058
4	5	400	0.31	-0.053	0.017	0.056	-0.048	0.015	0.050
4	6	200	0.31	-0.081	0.026	0.085	-0.076	0.025	0.080
4	6	400	0.31	-0.065	0.015	0.066	-0.059	0.012	0.060
8	5	200	0.23	-0.059	0.025	0.064	-0.052	0.023	0.057
8	5	400	0.23	-0.042	0.019	0.046	-0.040	0.015	0.043
8	5	200	0.23	-0.082	0.032	0.088	-0.086	0.026	0.084
8	6	400	0.23	-0.065	0.020	0.068	-0.061	0.017	0.063

Table 2. Simulation results for bias, standard deviation and root mean squared error for estimating $S(t_{0.5})$

				$\hat{S}_n(t_{0.5})$			$\hat{S}_M(t_{0.5})$		
θ	n_c	n	q_T	bias	std	rmse	bias	std	rmse
2	5	200	0.43	-0.072	0.025	0.076	-0.065	0.021	0.068
2	5	400	0.43	-0.056	0.014	0.057	-0.051	0.019	0.054
2	6	200	0.43	-0.080	0.026	0.084	-0.077	0.023	0.080
2	6	400	0.43	-0.068	0.019	0.071	-0.066	0.017	0.068
4	5	200	0.31	-0.058	0.025	0.063	-0.054	0.023	0.058
4	5	400	0.31	-0.047	0.015	0.050	-0.043	0.018	0.046
4	6	200	0.31	-0.079	0.035	0.086	-0.076	0.025	0.080
4	6	400	0.31	-0.061	0.017	0.063	-0.059	0.012	0.060
8	5	200	0.23	-0.057	0.025	0.062	-0.052	0.023	0.057
8	5	400	0.23	-0.037	0.019	0.042	-0.040	0.015	0.043
8	6	200	0.23	-0.078	0.029	0.083	-0.086	0.026	0.084
8	6	400	0.23	-0.064	0.018	0.066	-0.061	0.017	0.063

Table 3. Simulation results for bias, standard deviation and root mean squared error for estimating $S(t_{0.8})$

				$\hat{S}_n(t_{0.8})$			$\hat{S}_M(t_{0.8})$		
θ	n_c	n	q_T	bias	std	rmse	bias	std	rmse
2	5	200	0.43	-0.065	0.021	0.068	-0.061	0.025	0.066
2	5	400	0.43	-0.056	0.014	0.057	-0.052	0.018	0.055
2	6	200	0.43	-0.081	0.026	0.085	-0.077	0.023	0.080
2	6	400	0.43	-0.074	0.015	0.076	-0.072	0.013	0.073
4	5	200	0.31	-0.059	0.024	0.064	-0.054	0.023	0.058
4	5	400	0.31	-0.047	0.019	0.051	-0.048	0.015	0.050
4	6	200	0.31	-0.073	0.027	0.079	-0.076	0.025	0.080
4	6	400	0.31	-0.062	0.016	0.064	-0.059	0.012	0.060
8	5	200	0.23	-0.057	0.025	0.062	-0.052	0.023	0.057
8	5	400	0.23	-0.042	0.016	0.045	-0.040	0.015	0.043
8	6	200	0.23	-0.073	0.030	0.080	-0.086	0.026	0.084
8	6	400	0.23	-0.060	0.016	0.062	-0.061	0.017	0.063

4. Discussions

For interval-censored and left truncated data, Turnbull's algorithm leads to a self-consistent equation which is not in the form of an integral equation. Large sample properties of the NPMLE have not been previously examined because of, we believe, among other things, the lack of such an integral equation. In this article, we have presented a SCE using an integral equation and consistency of the SCE under some conditions (assumption (i) of Theorem 2). Since the NPMLE also satisfies the self consistent integral equation, the consistency of the NPMLE also holds. More research remains to be done. A rigorous investigation when assumption (i) of Theorem 2 holds. A similar equation holds for the simpler case: doubly censored data (see Gu and Zhang (1993)). Consider an alternative proof by extending the approach of Yu et al. (2001), where the consistency of SCE is established when there is no truncation.

References

Alioum A. and Commenges D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, **52**, 512-524.

Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **56**, 71-74.

Gentleman, R. and Geyer C. J. (1994). Maximum likelihood for interval censored data: consistency and computation. *Biometrika*, **81**, 618-623.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhäuser.

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.

Hudgens, M. G. (2005). On nonparametric maximum likelihood estimation with interval censoring and truncation. *Journal of the Royal Statistical Society, Series B*, **67**, part 4, 573-587.

Kalbfleish, J. D. and Lawless, J. F. (1989). Inferences based of retrospective ascertainment: An analysis of the data on transfusion related AIDS. *Journal of the American Statistical Association*, **84**, 360-372.

Li, L., Watkins, T., Yu, Q. Q. (1997). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. *Scand. J. Statist.*, **24**, 531-542.

Pan, W., Chappell, R. and Kosorok, M. R. (1998). On consistency of the monotone MLE of survival for left truncated and interval-censored data. *Statistics & Probability Letters.* **38**, 49-57.

Pan, W. and Chappell, R (1998). Computation of the NPMLE of distribution functions for interval censored and truncated data with applications to the Cox model. *Computational Statistics and Data Analysis*, **28**, 33-50.

Pan, W. and Chappell, R (1999). A note on inconsistency of NPMLE of the distribution function from left truncated and case I interval censored data. *Lifetime Data Analysis*, **5**, 281-291.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.*, **22**, 86-91.

Shen, P.-S. (2005). Estimation of the truncation probability with the left-truncated and right-censored data. *Nonparametric Statistics*, **17**, No. 8, 957-969.

Shick, A and Yu, Q. 2000. Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics*, **27**, 45-55.

Song, S. (2004). Estimation with univariate “mixed case” interval censored data. *Statist. Sin.*, **14**, 269-282.

Støvring, H. and Wang, M.-C. (2007). A new approach of nonparametric estimation of incidence and lifetime risk based on birth rates and incident events. *BMC Medical Research*, **7:53**, 1-11.

Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, **70**, 865-871.

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

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290-295.

van der Vaart, A. and Wellner, J. A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In High Dimensional Probability II, pp. 115-133. Boston: Birkhäuser.

Wang, M.-C. 1991, Nonparametric estimation from cross-sectional survival data. *J. Amer. Statist. Ass.*, *86*, 130-143.

Woodroofe, M., 1985, Estimating a distribution function with truncated data. *Ann. Statist.*,

13, 163-167.

Yu, Q., Li, L. and Wong, G.Y.C., (1998a). Asymptotic variance of the GMLE of a survival function with interval-censored data. *Sankhya*, **60**, 184-187.

Yu, Q., Shick, A., Li, L. and Wong, G.Y.C., (1998b). Asymptotic properties of the GMLE with case 2 interval-censored data. *Statistics & probability letters*, **37**, 223-228.

Yu, Q. Q., Li, L., and Wong, G. Y. C. (2000), On consistency of the self-consistent estimator of survival function with interval censored data. *Scan. J. of Statist.*, **27**, 35-44.

Yu, Q. Q., Wong, G. Y. C., and Li, L. (2001), Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Ann. Inst. Stat. Math.*, **53**, 469-486.