行政院國家科學委員會專題研究計畫 成果報告

區間設限及左截資料下之自我一致與非參數最大概似估計值
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 100-2118-M-029-002-
執 行 期 間 ：100 年 08 月 01 日至 101 年 07 月 31 日
執 行 單 位 ：東海大學統計學系

計畫主持人：沈葆聖

計畫參與人員：碩士班研究生-兼任助理人員：劉宜雅
碩士班研究生-兼任助理人員：石家瑜
碩士班研究生-兼任助理人員：劉怡芳
助教-兼任助理人員：蔡佳玲

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 101 年 09 月 03 日

中 文 摘 要 ： 區間設限意指存活時間 T 僅知落於某區間[L,R]。某些情形下，資料收集亦發生左截，即左截及區間設限資料。本研究依據積分方程式，我們將提出 T 的存活函數之自我一致估計值(SCE)，並證明非參數最大概似估計值(NPMLE)滿足自我一致積分方程式。經由模擬我們比較 SCE 和 NPMLE 之表現。我們亦檢討 SCE 之一致性。

中文關鍵詞： 左截, 區間設限, 自我一致性

英 文 摘 要 ：

英文關鍵詞：

☑ 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

區間設限及左截資料下之自我一致與非參數最大概似估計值

計畫類別：☑個別型計畫 □整合型計畫

計畫編號：NSC 100－2118－M－029－002－

執行期間：100 年 08 月 01 日至 101 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 101 年 07 月 29 日

Self-Consistent and Nonparametric Maximum Likelihood Estimators with Interval-Censored and Left-Truncated Data

Pao-sheng Shen

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

Abstract

Interval censoring refers to a situation in which, T_i^*, the time to occurrence of an event of interest is only known to lie in an interval $[L_i^*, R_i^*]$. In some cases, the variable T_i^* also suffers left-truncation. Based on an integral equation, we propose a self-consistent estimator (SCE) of survival function of T_i^*. It is shown that the nonparametric maximum likelihood estimator (NPMLE) is a solution of the integral equation. A simulation study is conducted to compare the performance between the SCE and NPMLE. We also discuss the consistency of the SCE.

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

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: 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 for each individual i the infection time (denoted by T_{si}) 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_{si} to development of AIDS. Let $V_i^* = \tau_0 - T_{si}$ if $T_{si} < \tau_0$ and $V_i^* = 0$ if $T_{si} \geq \tau_0$. Hence, T_i^* is observable only when $T_i^* \geq V_i^*$). Let $C_i^* = \tau_e - T_{si}$ 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_{si} is recorded exactly. Hence, the variable of interest T_i^* can be recorded as an interval, say $[L_i^*, R_i^*]$. For example, under mixed case interval-censored model (Shick and Yu (2000)), let K be a positive random integer, and for individual i let $\mathbf{Y_i} = \{Y_{i,k,j} : k = 1, 2 \ldots, j = 1, \ldots, k\}$ be an array of random variables such that $Y_{i,k,1} < \cdots < Y_{i,k,k}$. On the event $K = k$, let $[L_i^*, R_i^*]$ denote the endpoints of that random interval among $[-\infty, Y_{i,k,1}]$, $[Y_{i,k,1}, Y_{i,k,2}], \ldots, [Y_{k,k}, \infty]$ which contains T_i^*. When T_i^* is right censoring, we have $[L_i^*, R_i^*] = [Y_{k,k}, \infty) = [C_i^*, \infty]$. 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 $(K, \mathbf{Y_i}, V_i^*)$ and T_i^* are independent 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, for identifiability of T_i^*, we assume that T_i^*, L_i^*, R_i^* and V_i^* are all continuous, and

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

Furthermore, we assume that $P(L_i^* < R_i^*) = 1$ and given $R_i^* < \infty$, (L_i^*, R_i^*) has a joint density $b(l, r)$, satisfying $b(l, r) > 0$ if $0 < F(l) < F(r) < 1$.

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_i^*, R_i^*) is absolutely continuous. If (L_i^*, R_i^*) 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. We discuss 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 $\hat{S}_M = \text{argmax}_{S \in \mathcal{S}}\{L_c(S)\}$, 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 $j = 1, \ldots, J$. By Lemma 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 $q_1 \le p_1 < q_2 \le p_2 < \cdots < q_J \le p_J$. 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

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

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

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

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:

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

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 $\hat{s}_j$ $(j = 1, \ldots, J)$ 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 NPMLE 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 $\{1, \ldots, n\}$ 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 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 $0 < q_j < p_j < \infty$.

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:

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

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

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

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

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

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

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

$$\left. + \sum_{i=1}^{n} \frac{\hat{S}_n(q_j-) - \hat{S}_n(p_j)}{\hat{S}_n(V_i)} - \sum_{i=1}^{n} I_{[V_i \le q_j]} \frac{\hat{S}_n(q_j-)}{\hat{S}_n(V_i)} + \sum_{i=1}^{n} I_{[V_i \le p_j]} \frac{\hat{S}_n(p_j)}{\hat{S}_n(V_i)}\right\}. \quad (2.10)$$

Since there is no left truncation observations in $[q_j, p_j)$, (2.10) can be written as

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

$$\left. + \sum_{i=1}^n \frac{\hat{S}_n(q_j-) - \hat{S}_n(p_j)}{\hat{S}_n(V_i)} - \sum_{i=1}^n I_{[q_j \geq V_i]} \frac{\hat{S}_n(q_j-) - \hat{S}_n(p_j)}{\hat{S}_n(V_i)}\right\}. \tag{2.11}$$

Next,

$$\hat{S}_M(q_j-) - \hat{S}_M(p_j) = \left[\sum_{i=1}^n \frac{1}{\sum_{j=1}^J \beta_{ij} s_j}\right]^{-1} \left\{\sum_{i=1}^n \frac{\alpha_{ij}}{\sum_{k=1}^J \alpha_{ik}\hat{s}_k} + \sum_{i=1}^n \frac{1-\beta_{ij}}{\sum_{k=1}^J \beta_{ik}\hat{s}_k}\right\}\hat{s}_j. \tag{2.12}$$

By definitions of A_i, B_i, α_{ij} and β_{ij}, it follows that equation (2.11) is equivalent to equation (2.12). The proof is completed.

Although the NPMLE $\hat{S}_M$ satisfies equation (2.9), it is not clear whether the SCE is consistent or not. We discuss the consistency of the SCE as follows.

Let Ω be the event $\{\lim \tilde{H}_n(v,l) = \tilde{H}(v,l), \lim \tilde{Q}_n(l,r) = \tilde{Q}(l,r)$ 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: $S_0(t) =$

$$\left[\int \frac{1}{S_0(v)}\tilde{G}(dv)\right]^{-1}\left\{\int_{v\leq t<l} d\tilde{H}(v,l) + \int_{l\leq t<r} \frac{S_0(t) - S_0(r)}{S_0(l) - S_0(r)}\tilde{Q}(dl,dr) + \int_{v>t} \frac{S_0(t)}{S_0(v)}\tilde{G}(dv)\right\}. \tag{2.13}$$

Equation (2.13) can be written as

$$S_0(t)\int_{v\leq t} \frac{1}{S_0(v)}\tilde{G}(dv) = \int_{v\leq t<l} \tilde{H}(dv,dl) + \int_{l\leq t<r} \frac{S_0(t) - S_0(r)}{S_0(l) - S_0(r)}\tilde{Q}(dl,dr). \tag{2.14}$$

Let $H(v,l) = P(V_i^* \leq u, L_i^* \leq l)$ and $Q(l,r) = P(L_i^* \leq r, R_i^* \leq r)$. Since $\tilde{G}(dv) = p^{-1}S_F(v)G(dv)$, $\tilde{H}(dv,dl) = p^{-1}S_F(l)H(dv,dl)$, and $\tilde{Q}(dl,dr) = p^{-1}[S_F(l) - S_F(r)]Q(dl,dr)$, (2.14) can be written as

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

$$+p^{-1}\int_{l\leq t<r}\frac{S_0(t)-S_0(r)}{S_0(l)-S_0(r)}[S_F(l)-S_F(r)]Q(dv,dl,dr). \quad (2.15)$$

Replacing $S_0(\cdot)$ of (2.15) by $S_F(\cdot)$, we obtain

$$p^{-1}S_F(t)G(t)=p^{-1}\int_{v\leq t<l}S_F(l)H(dv,dl)+p^{-1}\int_{l\leq t<r}[S_F(t)-S_F(r)]Q(dl,dr). \quad (2.16)$$

Note that (2.16) is equivalent to

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

Subtracting (2.16) from (2.15), we obtain

$$h(t)K(t)=$$

$$S_0(t)\int_{v\leq t}\frac{h(v)}{S_0(v)}G(dv)-\int_{l\leq t<r}\frac{h(l)[S_0(t)-S_0(r)]}{S_0(l)-S_0(r)}Q(dl,dr)-\int_{l\leq t<r}\frac{h(r)[S_0(l)-S_0(t)]}{S_0(l)-S_0(r)}Q(dl,dr), \quad (2.17)$$

where $K(t)=G(t)-P(L_i^*\leq t<R_i^*)$ and $h(t)=S_0(t)-S_F(t)$. Hence, to obtain the consistency of the SCE, one need the following condition:

If (2.17) holds on $t\in(a_F,b_F)$ then $h(t)=0$ for $t\in(a_F,b_F)$ (2.18).

Hence, if (2.18) holds, we have $S_0(t)=S_F(t)$. Since all limit points of $\hat{S}_n$ must satisfy (2.15), by Helly-Bray selection theorem we have $\hat{S}_n(t)\rightarrow S_F(t)$ a.s. for $t\in(a_F,b_F)$ and $\sup_{t\in(a_F,b_F)}|\hat{S}_n(t)-S_F(t)|\rightarrow 0$ a.s if $\hat{S}_n$ is a sequence of monotone, right continuous and bounded functions on (a_F,b_F).

Similar to doubly censored data (see Gu and Zhang (1993)) condition (2.18) may hold if one can show that suppose $K(t)>0$ holds on $\{t:0<S(t)<1\}$ then $h(t)=0$ for all t provided that $h(t+)\neq h(t)\Rightarrow S(t+)<S(t)$ on $\{t:0<S(t)<1\}$, $h(t)=0$ on $\{t:S(t)=0 \text{ or } S(t)=1\}$. Although we are not able to establish the consistency of the SCE, the simulation study in Section 3 indicate that the SCE performs adequately.

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 $\theta=1,2$ and 4, i.e. $G(x;\theta)=1-\exp(-x\theta)$

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 $X_i = 2 + B(n_c, 0.5)$, where $B(n_c, 0.5)$ is a binomial random variable with $n_c = 5, 8$. Given $X_i = k$, 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. 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 $\hat{S}_M(t_p)$. For both estimators, the initial estimator is the product-limit estimator for left-truncated and right-censored data (see Wang (1987)) based on midpoint imputation. The convergence criterion was set $|\hat{S}_M^{(r+1)}(t_p) - \hat{S}_n^{(r)}(t_p)| < 0.001$ or $|\hat{S}_n^{(r+1)}(t_p) - \hat{S}_n^{(r)}(t_p)| < 0.001$. 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 root mean squared errors (rmse) 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) The biases of the estimators $\hat{S}_n$ are larger than that of $\hat{S}_M$ for most of the cases considered. In terms of rmse, 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 the 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
1	5	200	0.50	-0.016	0.029	0.033	-0.010	0.029	0.031
1	5	400	0.50	-0.009	0.020	0.022	-0.008	0.019	0.021
1	8	200	0.50	-0.014	0.029	0.032	-0.012	0.028	0.030
1	8	400	0.50	-0.006	0.018	0.020	-0.008	0.017	0.019
2	5	200	0.43	-0.012	0.039	0.041	-0.010	0.037	0.037
2	5	400	0.43	-0.009	0.021	0.022	-0.006	0.020	0.022
2	8	200	0.43	-0.013	0.036	0.038	-0.012	0.036	0.038
2	8	400	0.43	-0.007	0.021	0.022	-0.008	0.020	0.021
4	5	200	0.31	-0.011	0.037	0.039	-0.007	0.035	0.036
4	5	400	0.31	-0.005	0.021	0.022	-0.006	0.020	0.021
4	8	200	0.31	-0.015	0.035	0.040	-0.012	0.035	0.037
4	8	400	0.31	-0.007	0.019	0.020	-0.009	0.018	0.020

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
1	5	200	0.50	-0.008	0.059	0.060	-0.006	0.057	0.057
1	5	400	0.50	-0.010	0.025	0.027	-0.007	0.025	0.026
1	8	200	0.50	-0.012	0.057	0.058	-0.013	0.055	0.057
1	8	400	0.50	-0.008	0.022	0.023	-0.009	0.021	0.023
2	5	200	0.43	-0.023	0.055	0.060	-0.018	0.053	0.056
2	5	400	0.43	-0.012	0.038	0.041	-0.010	0.037	0.038
2	8	200	0.43	-0.026	0.053	0.059	-0.021	0.051	0.055
2	8	400	0.43	-0.014	0.036	0.038	-0.011	0.036	0.038
4	5	200	0.31	-0.036	0.067	0.076	-0.031	0.064	0.072
4	5	400	0.31	-0.016	0.040	0.041	-0.019	0.037	0.039
4	8	200	0.31	-0.031	0.064	0.071	-0.026	0.062	0.067
4	8	400	0.31	-0.013	0.037	0.039	-0.012	0.036	0.038

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
1	5	200	0.50	-0.012	0.048	0.049	-0.009	0.046	0.047
1	5	400	0.50	-0.008	0.027	0.028	-0.005	0.026	0.026
1	8	200	0.50	-0.010	0.045	0.046	-0.007	0.045	0.046
1	8	400	0.50	-0.009	0.026	0.027	-0.005	0.025	0.025
2	5	200	0.43	-0.032	0.069	0.076	-0.027	0.067	0.072
2	5	400	0.43	-0.017	0.041	0.044	-0.021	0.040	0.045
2	8	200	0.43	-0.029	0.066	0.072	-0.030	0.065	0.072
2	8	400	0.43	-0.016	0.039	0.042	-0.020	0.038	0.043
4	5	200	0.31	-0.041	0.073	0.083	-0.039	0.069	0.080
4	5	400	0.31	-0.023	0.048	0.053	-0.019	0.046	0.050
4	8	200	0.31	-0.032	0.070	0.077	-0.035	0.067	0.076
4	8	400	0.31	-0.021	0.046	0.051	-0.020	0.045	0.049

4. Applications

For purpose of illustration, we apply both estimators to the CDC AIDS Blood Transfusion Data. The AIDS Blood Transfusion Data are collected by the Centers for Disease Control (CDC), which is from a registry data base, a common source of medical data (see Kalbfleish and Lawless (1989)). The data were retrospectively ascertained for all transfusion-associated AIDS cases in which the diagnosis of AIDS occurred prior to the end of of the study, which was June 30, 1991. The data consist of the time in month and only cases having either one transfusion or multiple transfusions in the same calendar month were used. Cases having the AIDS prior to July 1, 1982 (τ_0) were not included because this is when adults started being infected by the virus from a contaminated blood transfusion. Because HIV was unknown prior to 1982, and cases of transfusion-related AIDS before τ_0 would have been missed (i.e. left-truncated). Let T_{si} be the calendar time (in years) of the initiating events (HIV infection), and τ_D be the calendar time (in years) at which AIDS is diagnosed. Let $T_i^* = 12(\tau_D - T_{si})$ (in month) be the incubation time from HIV infection to AIDS. Let $V_i^* = 12(\tau_0 - T_{si})$ (in month) denote the left-truncated variable. Hence, T_i^* is observable only when $T_i^* \geq V_i^*$). There were 295 truncated observations. To introduce interval censoring, similar to the setup in simulation study, we generate a random variable $X = 2 + B(6, 0.8)$. Given $X_i = k$, we then generate k i.i.d uniform random variables $U_{ji} \sim U(0, 1)$ $(j = 1, \ldots, k)$. Using the approach of Section 3, we obtain the truncated interval observations (L_i, R_i, V_i) $(i =$

$1, \ldots, 295$). For purpose of comparison we also obtain the estimators of $S(t)$ (denoted by $\hat{S}_E$) by using the exact observations (T_i, V_i)'s (i.e. left-truncated data). Table 4 shows the results of the three estimators $\hat{S}_M$, $\hat{S}_n$ and $\hat{S}_E$ at some selected values of t. Table 4 indicates that the differences between $\hat{S}_M(t)$ and $\hat{S}_E(t)$ (denoted by diff1) are smaller than that between $\hat{S}_n(t)$ and $\hat{S}_E(t)$ (denoted by diff2).

Table 4. Estimation of the distribution function
of the incubation time for AIDS Blood Transfusion Data

t	$\hat{S}_E(t)$	$\hat{S}_M(t)$	$\hat{S}_n(t)$	diff1	diff2
10	0.835	0.819	0.813	-0.016	-0.022
15	0.684	0.669	0.665	-0.015	-0.019
20	0.577	0.565	0.559	-0.012	-0.016
25	0.456	0.447	0.443	-0.009	-0.013
30	0.372	0.359	0.354	-0.013	-0.018
35	0.287	0.276	0.271	-0.011	-0.016
40	0.204	0.193	0.188	-0.011	-0.015
45	0.158	0.150	0.148	-0.008	-0.010
50	0.114	0.104	0.101	-0.010	-0.013
55	0.102	0.094	0.090	-0.008	-0.012
60	0.091	0.085	0.082	-0.006	-0.009
70	0.070	0.066	0.062	-0.004	-0.008
80	0.052	0.047	0.046	-0.005	-0.006

5. 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 shown that the NPMLE is a solution of the integral equation. If we can show the consistency of the SCE under certain conditions then the consistency of the NPMLE can therefore be established. More research is needed to investigate this problem.

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. *Statistica Sinica*, **14**, 269-282.

Song, S. (2004). Estimation with univariate “mixed case” interval censored data. *Statistica Sinica*, **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. (1987). Product-limit estimates: a generalized maximum likelihood study. *Communications in Statistics, Part A- Theory and Methods*, **6**, 3117-3132.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, *86*, 130-143.

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

國科會補助計畫衍生研發成果推廣資料表

日期:2012/08/21

國科會補助計畫	計畫名稱: 區間設限及左截資料下之自我一致與非參數最大概似估計值
	計畫主持人: 沈葆聖
	計畫編號: 100-2118-M-029-002- 學門領域: 存活(倖存)分析

無研發成果推廣資料

100 年度專題研究計畫研究成果彙整表

計畫主持人：沈葆聖			計畫編號：100-2118-M-029-002-				
計畫名稱：區間設限及左截資料下之自我一致與非參數最大概似估計值							
成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	3	10%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 ■未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

投稿中

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

有助於對 T 區間設限及左截資料下之無母數最大概似估計值之漸近性質之推導