東海大學統計學研究所

碩士論文

指導教授：沈葆聖　教授

A General Semiparametric Model for Left-Truncated and Right-Censored Data

研究生：林文琦

中華民國九十五年七月

A General Semiparametric Model for Left-Truncated and Right-Censored Data

Wen-Chi Lin
Dept. of Statistics
Tunghai University
Taichung, 40704
Taiwan, R.O.C.

July 4, 2006

謝誌

本篇論文能夠順利完成，必須感謝我的指導教授沈葆聖博士。在我研究所的這二年中，老師非常用心的教導，無論在課業，或是在做人處事的道理上，老師都常常提供我最大的幫助。而老師對於做研究的積極與努力不懈的態度，更是讓我深受影響，並常以他做為最好的學習榜樣。感謝老師在這二年中對我的照顧與關心，您的付出，我深深感激。

其次，我要感謝黃連成教授以及戴政教授，在口試時提供了許多寶貴的意見，使我獲益匪淺；還要感謝東海統研所的教授們，在這二年中對我的教導與鼓勵，因為你們，才使我學到了這麼多的學問，也使我在研究所的這二年中，感覺非常踏實，生活的很有意義。

另外，要感謝研究室的學長姊、學弟妹，以及我的同學們，因為你們的陪伴，使我感覺我並不孤單。謝謝你們在學業上對我的幫助，以及在生活上對我的照顧，讓我這個外地人，有了第二個家的感覺，也給了我許多許多美好的回憶。

最後，我要感謝我的親愛的家人，一直在我求學的過程中，給我最大的支持與鼓勵；在我遇到挫折時，不斷的給我力量及依靠，使我能克服困難，順利的完成我的求學歷程。

感謝所有對我好的人，因為你們，使我擁有了正面積極的態度，也讓我能以微笑，來面對我的人生。謝謝你們！

Content

List of Tables

Abstract

In many follow-up studies survival data are often observed according to a cross-sectional sampling scheme. Data of this type are subject to left truncation and right censoring. In many practical cases, two types of censoring may occur. The first type of censoring (type A) is due to termination of the follow-up period. The second type of censoring (type B) is a consequence of other types of failure which might occur before the cross-section time. Let T^*, V^*, C_1^* and C_2^* denote the lifetime, left truncation, type A and type B censoring variables, respectively. Assume that T^*, (V^*, C_1^*) and C_2^* are independent of one another but V^* and C_1^* are dependent with $P(C_1^* \geq V^*) = 1$. Let F, G and Q denote the common distribution functions of T^*, V^* and C_2^*, respectively. Let $Z^* = \min(T^*, C_2^*)$. For left-truncated and right-censored (LTRC) data, one can observe nothing if $Z^* < V^*$, and observe (X^*, δ^*), if $Z^* \geq V^*$, where $X^* = \min(Z^*, C_1^*)$, and δ^* is equal to one if $X^* = T^*$, equal to two if $X^* = C_1^*$ and zero otherwise. For LTRC data, the truncation product-limit estimate $\hat{F}_n$ is the maximum likelihood estimate (MLE) for nonparametric models. If the distribution of V^* is parameterized as $G(x; \theta)$ and the distributions of T^* and C_2^* are left unspecified, the product-limit estimate $\hat{F}_n$ is not the MLE for this semiparametric model. When $C_1^* = C_2^* = \infty$ (i.e. left-truncated data), Wang (1989) derived the MLE of F for the semiparametric model and established its weak convergence properties. When $G(x; \theta) = x/\theta$ and $C_2^* = \infty$ (the so-called stationarity assumption), Asgharian et al. (2002, 2005) derived an unconditional MLE of F and established its asymptotic properties. In this note, we extend previous models by distinguishing two types of censoring. Iterative algorithms are proposed to obtain a semiparametric estimate, $\hat{F}_n(x; \hat{\theta}_n)$. The consistency of $\hat{F}_n(x; \hat{\theta}_n)$ is established. A simulation study is conducted to compare the performance of $\hat{F}_n(x; \hat{\theta}_n)$ against that of $\hat{F}_n(x)$.

Key Words: Left truncation, right censoring, conditional likelihood.

1. Introduction

In many follow-up studies involving cross-sectional sampling, an individual is observed only when a certain sampling status is satisfied. Data of this type are subject to left truncation and right censoring (see Wang (1991) for further details). In some case, the censoring (type A) is restricted to termination of the follow-up period. However, in many practical cases, censoring (type B) is a consequence of other types of failure which might occur before the cross-section time. Let T^*, V^*, C_1^* and C_2^* denote the lifetime, left truncation, type A and type B censoring variables, respectively. Assume that T^*, (V^*, C_1^*) and C_2^* are independent of one another but V^* and C_1^* are dependent with $P(C_1^* \geq V^*) = 1$. Let F, G and Q denote the common distribution functions of T^*, V^* and C_2^*, respectively. Let $Z^* = \min(T^*, C_2^*)$. For left-truncated and right-censored (LTRC) data, one can observe nothing if $Z^* < V^*$, and observe (X^*, δ^*) if $Z^* \geq V^*$, where $X^* = \min(Z^*, C_1^*)$ and δ^* is equal to one if $X^* = T^*$, equal to two if $X^* = C_1^*$ and zero otherwise. Consider the following examples.

Example 1.1 (Channing House data)

Channing House is a retirement center in Palo Alto, California. The data were collected between the opening of the house in January 1964 and July 1, 1975. In that time 97 men and 365 women passed through the center. Some of the individuals were censored due to leaving. The left truncation variable (V^*) here is the entry age into the Channing House and type B censoring (C_2^*) variable is the age on leaving. It is clear that only subjects with entry age (V^*) smaller than or equal to age on leaving (C_2^*) and death (T^*), i.e. $Z^* \geq V^*$, can become part of the sample. Moreover, a large number of the observations were censored due to the residents being alive on July

1, 1975 (termination of the follow-up). Hence, type A censoring variable (C_1^*) is the censored age on July, 1975, and the relationship $C_1^* \geq V^*$ is always satisfied.

Example 1.2 (Life-testing data)

Assume that n objects were put in use at some time in the distant past. These objects may fail due to type 1 or type 2 causes; whenever an object failed (type 1 or type 2), it was promptly replaced by another member of the same population. The parameters of interest are the distribution functions, F and Q, of the lifetimes for these objects until failure type 1 (T^*) or type 2 (C_2^*). At some time t_0 long after the start of the process a statistician arrives on the scene. It is assumed that the age (V^*) of each object in use at t_0 is known. Hence, his observation is restricted to the n objects in use at that time, i.e. $Z^* \geq V^*$. Suppose that each object is observed until $t_0 + d_0$. Hence, type A censoring variable ($C_1^* = V^* + d_0$) is induced due to termination of the follow-up period.

Let a_F and b_F denote the left and right endpoints of F. Define (a_G, b_G) and (a_Q, b_Q) similarly. For identifiablities of F, G,and Q, we assume that

$$a_G \leq \min(a_F, a_Q) \; and \; b_G \leq \min(b_F, b_Q). \tag{1.1}$$

Let $(X_1, \delta_1, V_1), \ldots, (X_n, \delta_n, V_n)$ denote the left-truncated and right-censored sample.

Let $R_n(u) = n^{-1} \sum_{i=1}^n I_{[V_i \leq u \leq X_i]}$, $N_F(u) = \sum_{i=1}^n I_{[X_i \leq u, \delta_i = 1]}$ and $N_F(du) = N_F(u) - N_F(u-)$. Suppose that $nR_n(X_i) > N_F(dX_i)$ for $i = 1, \ldots, n$ (see Wang (1987)). Then, the nonparametric maximum likelihood estimate (NPMLE) of $F(x)$ is given by

$$\hat{F}_n(x) = 1 - \prod_{u \leq x} \Big[1 - \frac{N_F(du)}{nR_n(u)}\Big].$$

Note that when G is not specified, the NPMLE of F is obtained condition on the observed V_i's. There are many applications (such as example 1.2), however, in which the initiation times follow a stationary Poisson process which implies $G(x;\theta) = x/\theta$ (the so-called stationarity assumption or length-biased sampling). When $G(x;\theta) = x/\theta$ and $C_2^* = \infty$, Wang (1991) had suggested that an unconditional likelihood approach is more efficient than its conditional counterpart. This improvement in efficiency was later confirmed by Asgharian, M'Lan and Wolfson (2002). Under the model of stationarity and $C_2^* = \infty$, Asgharian and Wolfson (2005) established the asymptotic properties of the unconditional MLE of F. A compromise between the stationarity assumption and the nonoparametric assumption on G would be the parameterized $G(x;\theta)$, where $\theta \in \Theta \subset R^q$, and θ is a q-dimensional vector. In example 1.1, the truncation distribution G can be interpreted as the distribution of the potential elderly resident's age at entry, which can follow uniform or some other distributions. In Section 2, we consider a general semiparametric model by distinguishing two types of censoring. Iterative algorithms are proposed to obtain a semiparametric estimate, $\hat{F}_n(x;\hat{\theta}_n)$. The consistency of $\hat{F}_n(x;\hat{\theta}_n)$ is established. In Section 3, a simulation study is conducted to compare the performance of $\hat{F}_n(x;\hat{\theta}_n)$ against that of $\hat{F}_n(x)$.

2. Semiparametric Estimates

2.1 Notations

Let $T_1, \ldots, T_{n_D}$ be the observed failure times, i.e. the observations from the subset $\mathcal{D} = \{X_i : \delta_i = 1; i = 1, \ldots, n\}$, $C_{11}, \ldots, C_{1n_A}$ be the observed type A censoring times, i.e. the observations from the subset $\mathcal{C}_\mathcal{A} = \{X_i : \delta_i = 2; i = 1, \ldots, n\}$, and $C_{21}, \ldots, C_{2n_B}$ be the observed type B censoring times, i.e. the observations from the subset $\mathcal{C}_\mathcal{B} = \{X_i : \delta_i = 0; i = 1, \ldots, n\}$. Let $Z_1, \ldots, Z_{n_K}$ be the observations from the subset $\mathcal{D} \cup \mathcal{C}_\mathcal{B}$. Let $x_1 < \cdots < x_m$ denote the distinct values of $Z_1, \ldots, Z_{n_K}$ and $C_{11}, \ldots, C_{1n_A}$ in increasing order.

For $j = 1, \ldots, m$, let $t_j = \sum_{i=1}^{n_D} I_{[T_i = x_j]}$, $c_{1j} = \sum_{i=1}^{n_A} I_{[C_{1i} = x_j]}$, $c_{2j} = \sum_{i=1}^{n_B} I_{[C_{2i} = x_j]}$, and $k_j = t_j + c_{2j}$.

Let $K(x)$ denote the distribution function of Z^* and $\bar{K}(x) = (1 - F(x))(1 - Q(x))$.

2.2 Estimation of θ

First, we consider the estimation of $K(x)$. Given θ, the marginal likelihood of Z_i's is given by

$$L(K; \theta) = \prod_{j=1}^{m} \Big(\frac{K(dx_j)}{\alpha} \Big)^{k_j} \prod_{j=1}^{m} \Big(\frac{\bar{K}(x_j)}{\alpha} \Big)^{c_{1j}},$$

where $\alpha = \int G(x; \theta) K(dx)$ and $K(dx_j) = K(x_j) - K(x_j -)$.

For a fixed θ, maximizing $L(K; \theta)$ with respect to $K(dx_j)$ is equivalent to maximizing

$$L^*(K; \theta) = \prod_{j=1}^{m} \Big(\frac{G(x_j; \theta) K(dx_j)}{\alpha} \Big)^{k_j} \prod_{j=1}^{m} \Big(\frac{\bar{K}(x_j)}{\alpha} \Big)^{c_{1j}},$$

subject to $K(dx_j) \geq 0$ and $\sum_{j=1}^{m} K(dx_j) = 1$.

Let $H(dx;\theta) = G(x;\theta)K(dx)/\alpha$. Then the problem of maximizing $L^*(K;\theta)$ is equivalent to that of maximizing

$$L(H;\theta) = \prod_{j=1}^{m}[H(dx_j;\theta)]^{k_j} \prod_{j=1}^{m}\Big(\int_{z \geq x_j} \frac{1}{G(z;\theta)} H(dz;\theta)\Big)^{c_{1j}}.$$

Note that when $G(x;\theta) = x/\theta$, the likelihood $L(H;\theta)$ is reduced to the likelihood for problem A of Vardi (1989). In considering the problem of estimating survivor function from multiplicatively right censored data, Vardi (1989) derived the unconditional NPMLE of a length-biased survival function from informatively censored data. The following example extend Vardi's problem A (see Vardi (1989), page 751).

Example 2.1: Multiplicative censoring

Let $W_1, \ldots, W_{n_K}$ and $W_1^c, \ldots, W_{n_A}^c$ be i.i.d. random variables from the lifetime distribution function $H(x;\theta)$, let $U_1, \ldots, U_{n_A}$ be i.i.d. uniform (0,1) random variables, and write $Y_i = G_\theta^{-1}(G(W_i^c;\theta)U_i)$, where $G_\theta^{-1}(z)$ denote the inverse function of $G(z;\theta)$. Given θ, we want to derive the nonparametric MLE of $H(x;\theta)$ based on the data $W_1, \ldots, W_{n_K}$ and $Y_1, \ldots, Y_{n_A}$.

Since

$$\begin{aligned} H_A(y;\theta) &= P(Y_i \leq y) = P(G(W_i^c;\theta)U_i \leq G(y;\theta)) \\ &= \int_{z \geq y} \frac{G(y;\theta)}{G(z;\theta)} H(dz;\theta) + H(y;\theta). \end{aligned}$$

Assume that $G(x;\theta)$ has a density function $g(x;\theta)$. Hence, the probability density function of Y_i is given by

$$h_A(y;\theta) = \int_{z \geq y} \frac{g(y;\theta)}{G(z;\theta)} H(dz;\theta) \quad y > 0.$$

Therefore, the marginal likelihood of W_i's and Y_i's is given by

$$L^*(H;\theta) = \prod_{i=1}^{n_K}[H(dW_i;\theta)] \prod_{i=1}^{n_A}\Big(\int_{z\geq Y_i} \frac{g(Y_i;\theta)}{G(z;\theta)} H(dz;\theta)\Big).$$

For a fixed θ, the likelihood function $L^*(H;\theta)$ treats $g(Y_i;\theta)$ as constants. Hence, the likelihood function $L^*(H;\theta)$ is equivalent to $L(H;\theta)$ by writing $W_i = Z_i$ ($i = 1,\ldots,n_K$) and $Y_i = C_{1i}$ ($i = 1,\ldots,n_A$).

For $j = 1,\ldots,m$, let $p_j = H(dx_j;\theta)$. The problem of maximizing $L(H;\theta)$ is reduced to maximizing

$$L(p;\theta) = \prod_{j=1}^{m} p_j^{k_j}\Big(\sum_{k=j}^{m} \frac{1}{G(x_k;\theta)} p_k\Big)^{c_{1j}},$$

subject to $p_j \geq 0$ ($j = 1,\ldots,m$) and $\sum_{j=1}^{m} p_j = 1$. Similar to Vardi's (1989) approach, the following EM algorithm is used to find out the MLEs of p_j's.

Initialization: Start with an arbitrary $p^{old} = [p_1^{old},\ldots,p_m^{old}]$ satisfying for $j = 1,\ldots,m$, $p_j^{old} > 0$ and $\sum_{j=1}^{m} p_j^{old} = 1$.

Iteration step: Replace p_j^{old} with

$$\begin{aligned} p_j^{new} &= n^{-1}E\Big[\sum_{i=1}^{n_K} I_{[Z_i=x_j]} + \sum_{i=1}^{n_A} I_{[C_{1i}=x_j]}\Big| Z_1,\ldots,Z_{n_K}, C_{11},\ldots,C_{1n_A}, p^{old}\Big] \\ &= n^{-1}\Big[k_j + \frac{1}{G(x_j;\theta)} p_j^{old} \sum_{k=1}^{j} c_{1k}\Big(\sum_{i=k}^{m} \frac{1}{G(x_i;\theta)} p_i^{old}\Big)^{-1}\Big]. \end{aligned} \tag{2.1}$$

Given θ, it follows that there exists a unique maximizer, $\hat{p}(\cdot;\theta)$ of the likelihood function $L(p;\theta)$ (see Vardi (1989), page 755). Let $\hat{H}_n(dx_j;\theta) = \hat{p}(x_j;\theta)$. Given $\hat{H}_n(dx;\theta)$, we can obtain the maximizer of $L(K;\theta)$, $\hat{K}_n(dx;\theta)$ by

$$\hat{K}_n(dx;\theta) = \frac{[G(x;\theta)]^{-1}\hat{H}_n(dx;\theta)}{\int_0^\infty [G(u;\theta)]^{-1}\hat{H}_n(du;\theta)}. \tag{2.2}$$

Based on (2.1) and (2.2), we can estimate θ using the following iterative algorithm.

For fixed $K(dx_1), \ldots, K(dx_m)$, the marginal likelihood of $V_1, \ldots, V_n$ is given by

$$L_v(\theta; K(dx)) = \prod_{i=1}^{n} \frac{g(V_i; \theta) \sum_{j=1}^{m} I_{[x_j \geq V_i]} K(dx_j)}{\sum_{j=1}^{m} G(x_j; \theta) K(dx_j)}.$$

Since the likelihood $L_v(\theta; K(dt))$ treats $I_{[x_j \geq V_i]} K(dx_j)$ as constants, the log-likelihood is

$$\log L_v(\theta; K(dx)) = \sum_{i=1}^{n} \log g(V_i; \theta) - n \log\Big(\sum_{j=1}^{m} G(x_j; \theta) K(dx_j)\Big).$$

For $j = 1, \ldots, m$, we use the product-limit estimate, $\hat{K}_n^{(0)}(dx) = \hat{K}_n^{(0)}(x) - \hat{K}_n^{(0)}(x-)$, as the initial estimator, where

$$\hat{K}_n^{(0)}(x) = 1 - \prod_{u \leq x} \Big[1 - \frac{N_K(du)}{nR_n(u)}\Big],$$

where $N_K(u) = \sum_{i=1}^{n_K} I_{[Z_i \leq u]}$.

Step 1: For fixed $\hat{K}_n^{(0)}(dx_1), \ldots, \hat{K}_n^{(0)}(dx_m)$, maximize $L_v(\theta; \hat{K}_n^{(0)}(dx))$ with respect to θ. Let $\hat{\theta}^{(1)}$ denote the unique maximizer of $L_v(\theta; \hat{K}_n^{(0)}(dx))$.

Step 2: For fixed $\hat{\theta}^{(1)}$, a unique maximizer $\hat{p}(\cdot; \hat{\theta}^{(1)})$ of the likelihood function $L(p; \hat{\theta}^{(1)})$ can be obtained by (2.1). Given $\hat{H}_n(dx; \hat{\theta}^{(1)})$, we can obtain the maximizer of $L(K; \hat{\theta}^{(1)})$, $\hat{K}_n^{(1)}(dx; \hat{\theta}^{(1)})$ by

$$\hat{K}_n^{(1)}(dx; \hat{\theta}^{(1)}) = \frac{[G(x; \hat{\theta}^{(1)})]^{-1} \hat{H}_n(dx; \hat{\theta}^{(1)})}{\int_0^\infty [G(u; \hat{\theta}^{(1)})]^{-1} \hat{H}_n(du; \hat{\theta}^{(1)})}.$$

Repeat steps 1 and 2 until the solution is stable. Let $\hat{\theta}_n$, $\hat{H}_n(dx_j; \hat{\theta}_n)$'s and $\hat{K}_n(dx_j; \hat{\theta}_n)$'s denote the stable solutions. Let $\hat{H}_n(x; \hat{\theta}_n) = \sum_{j=1}^{m} \hat{H}_n(dx_j; \hat{\theta}_n) I_{[x_j \leq x]}$. Define $\hat{K}_n(x; \hat{\theta}_n)$ similarly.

Assume that $K(x)$ has a density function $k(x)$. Let

$$\log L_v(\theta; k(x)) = \sum_{i=1}^{n} \log g(V_i; \theta) - n \log \int_{a_K}^{b_K} G(x; \theta) k(x) dx.$$

Since the product-limit estimator $\hat{K}_n^{(0)}(x)$ is uniformly consistent, we have

$$|\log L_v(\theta; \hat{K}_n^{(0)}(dx)) - \log L_v(\theta; k(x))| \to 0$$

as $n \to \infty$. Hence, the strong consistence of $\hat{\theta}_n$ can be established. Similar to the proof of Theorem 3.1 of Wang (1989), we need the following assumptions to derive the consistency of $\hat{H}_n(x; \hat{\theta}_n)$ and $\hat{K}_n(x; \hat{\theta}_n)$:

(a) K is continuous.

(b) $G(x; \theta)$ is continuous in x for each $\theta \in \Theta$.

(c) $\hat{\theta}_n \xrightarrow{p} \theta$ implies $G(x; \hat{\theta}_n) \to G(x; \theta)$ for each x.

Lemma 2.1

Under assumptions (a), (b) and (c), if $n_K/(n_K + n_A) \to p_K > 0$ then $\sup_{a_K \le x \le b_K} |\hat{H}_n(x; \hat{\theta}_n) - H(x; \theta)| \to 0$ with probability 1.

Proof : The proof is technical and is omitted.

2.3 Estimation of $F(x)$

Given $\hat{\theta}_n$, the estimated marginal likelihood of T_i's and C_{2i}'s is given by

$$L(F, Q; \hat{\theta}_n) = \prod_{j=1}^{m} \Big(\frac{F(dx_j)\bar{Q}(x_j-)}{\hat{\alpha}}\Big)^{t_j} \prod_{j=1}^{m} \Big(\frac{Q(dx_j)\bar{F}(x_j-)}{\hat{\alpha}}\Big)^{c_{2j}} \prod_{j=1}^{m} \Big(\frac{\bar{F}(x_j-)\bar{Q}(x_j-)}{\hat{\alpha}}\Big)^{c_{1j}},$$

where $\hat{\alpha} = \int G(x; \hat{\theta}_n) K(dx)$.

Let $G(x_j;\hat{\theta}_n)F(dx_j)\bar{Q}(x_j-)/\hat{\alpha} = \tilde{F}(dx_j)$ and $G(x_j;\hat{\theta}_n)Q(dx_j)\bar{F}(x_j-)/\hat{\alpha} = \tilde{Q}(dx_j)$. Then $L(F,Q;\hat{\theta}_n)$ can be written as

$$L(\tilde{F},\tilde{Q};\hat{\theta}_n) = \prod_{j=1}^{m}\Big(\tilde{F}(dx_j)\Big)^{t_j}\prod_{j=1}^{m}\Big(\tilde{Q}(dx_j)\Big)^{c_{2j}}\prod_{j=1}^{m}\Big(\sum_{k=j}^{m}\frac{1}{G(x_k;\hat{\theta}_n)}[\tilde{F}(dx_k)+\tilde{Q}(dx_k)]\Big)^{c_{1j}}.$$

For $j = 1,\dots,m$, let $\tilde{p}_j = \tilde{F}(dx_j)$ and $\tilde{q}_j = \tilde{Q}(dx_j)$. The problem of maximizing $L(\tilde{F},\tilde{Q};\hat{\theta}_n)$ is reduced to maximizing

$$L(\tilde{p},\tilde{q};\hat{\theta}_n) = \prod_{j=1}^{m}\tilde{p}_j^{t_j}\tilde{q}_j^{c_{2j}}\Big(\sum_{k=j}^{m}\frac{1}{G(t_k;\theta)}(\tilde{p}_k+\tilde{q}_k)\Big)^{c_{1j}},$$

subject to $\tilde{p}_j \geq 0, \tilde{q}_j \geq 0$ $(j = 1,\dots,m)$ and $\sum_{j=1}^{m}(\tilde{p}_j+\tilde{q}_j) = 1$. Similar to Vardi's (1989) approach, the following EM algorithm is used to find out the MLEs of $\tilde{p}_j$'s and $\tilde{q}_j$.

Initialization: Start with an arbitrary $\tilde{p}^{old} = [\tilde{p}_1^{old},\dots,\tilde{p}_m^{old}]$ and $\tilde{q}^{old} = [\tilde{q}_1^{old},\dots,\tilde{q}_m^{old}]$ satisfying for $j = 1,\dots,m$, $\tilde{p}_j^{old} > 0$, $\tilde{q}_j^{old} > 0$ and $\sum_{j=1}^{m}(\tilde{p}_j^{old}+\tilde{q}_j^{old}) = 1$.

Iteration step: Replace $\tilde{p}_j^{old}$ and $\tilde{q}_j^{old}$ with

$$\tilde{p}_j^{new} = n^{-1}E\Big[\sum_{i=1}^{n_D}I_{[T_i=x_j]}+\sum_{i=1}^{n_A}I_{[C_{1i}=x_j]}\Big|T_1,\dots,T_{n_D},C_{21},\dots,C_{2n_B},C_{11},\dots,C_{1n_A},\tilde{p}^{old},\tilde{q}^{old}\Big]$$

$$= n^{-1}\Big[t_j+\frac{1}{G(x_j;\hat{\theta}_n)}\tilde{p}_j^{old}\sum_{k=1}^{j}c_{1k}\Big(\sum_{i=k}^{m}\frac{1}{G(x_i;\hat{\theta}_n)}(\tilde{p}_i^{old}+\tilde{q}_i^{old})\Big)^{-1}\Big]$$

and

$$\tilde{q}_j^{new} = n^{-1}E\Big[\sum_{i=1}^{n_B}I_{[C_{2i}=x_j]}+\sum_{i=1}^{n_A}I_{[C_{1i}=x_j]}\Big|T_1,\dots,T_{n_A},C_{21},\dots,C_{2n_B},C_{11},\dots,C_{1n_A},\tilde{p}^{old},\tilde{q}^{old}\Big]$$

$$= n^{-1}\Big[c_{2j}+\frac{1}{G(x_j;\hat{\theta}_n)}\tilde{q}_j^{old}\sum_{k=1}^{j}c_{1k}\Big(\sum_{i=k}^{m}\frac{1}{G(x_i;\hat{\theta}_n)}(\tilde{p}_i^{old}+\tilde{q}_i^{old})\Big)^{-1}\Big].$$

Let $\tilde{F}_n(dx_j;\hat{\theta}_n)$ and $\tilde{Q}_n(dx_j;\hat{\theta}_n)$ denote the maximizer of $L(\tilde{F},\tilde{Q};\hat{\theta}_n)$. Based on $\tilde{F}_n(dx_j;\hat{\theta}_n)$ and $\tilde{Q}_n(dx_j;\hat{\theta}_n)$, a semiparametric estimator of F is given by

$$\hat{F}_n(x;\hat{\theta}_n) = 1 - \prod_{x_j \le x}\Big[1 - \frac{\tilde{F}_n(dx_j;\hat{\theta}_n)}{G(x_j;\hat{\theta})\tilde{C}_n(x_j;\hat{\theta}_n)}\Big],$$

where $\tilde{C}_n(x_j;\hat{\theta}_n) = \sum_{k\ge j}\frac{1}{G(x_k;\hat{\theta}_n)}\big(\tilde{F}_n(dx_k;\hat{\theta}_n) + \tilde{Q}_n(dx_k;\hat{\theta}_n)\big)$.

The $\tilde{F}_n(dt;\hat{\theta}_n)$ and $\tilde{Q}_n(dt;\hat{\theta}_n)$ must satisfy the following two score equations

$$\tilde{F}_n(dx;\hat{\theta}_n) = \frac{n_D}{n}\hat{H}_{n_D}(dx) + \frac{n_A}{n}\int_{0<y\le x}\frac{\hat{H}_{n_A}(dy)}{\int_{z\ge y}[G(z;\hat{\theta}_n)]^{-1}\tilde{H}_n(dz;\hat{\theta}_n)}\frac{1}{G(x;\hat{\theta}_n)}\tilde{F}_n(dx;\hat{\theta}_n), \tag{2.3}$$

$$\tilde{Q}_n(dx;\hat{\theta}_n) = \frac{n_B}{n}\hat{H}_{n_B}(dt) + \frac{n_A}{n}\int_{0<y\le x}\frac{\hat{H}_{n_A}(dy)}{\int_{z\ge y}[G(z;\hat{\theta}_n)]^{-1}\tilde{H}_n(dz;\hat{\theta}_n)}\frac{1}{G(x;\hat{\theta}_n)}\tilde{Q}_n(dx;\hat{\theta}_n), \tag{2.4}$$

subject to $\sum_{j=1}^m \tilde{H}_n(dx_j;\hat{\theta}_n) = 1$, where $\tilde{H}_n(dz;\hat{\theta}_n) = \tilde{F}_n(dz;\hat{\theta}_n) + \tilde{Q}_n(dz;\hat{\theta}_n)$, $\hat{H}_{n_D}$ and $\hat{H}_{n_B}$ denote the empirical distribution function of T_i's and C_{2i}'s, respectively. By Lemma 2.1, (2.3) and (2.4), it follows that $\tilde{H}_n(dx;\hat{\theta}_n) = \hat{H}_n(dx;\hat{\theta}_n)$. By Lemma 2.1, $\tilde{H}_n(dx;\hat{\theta}_n)$ is a consistent estimator of $H(dx;\theta)$.

Next, we derive the consistency of $\tilde{F}_n(x;\hat{\theta}_n)$ and $\tilde{Q}_n(x;\hat{\theta}_n)$. First, $(n_D/n)\hat{H}_{n_D}(dx)$ is a consistent estimator of $\alpha^{-1}F(dx)\bar{Q}(x)P(V^* \le x \le C_1^*)$. Similarly, $(n_A/n)\hat{H}_{n_A}(dy)$ is a consistent estimator of $\alpha^{-1}A(dy)\bar{K}(y)$, where $A(dy) = A(y) - A(y-)$, $A(y) = P(C_1^* \le y)$. By Lemma 2.1, it follows that $\int_{z\ge y}[G(z;\hat{\theta}_n)]^{-1}\tilde{H}_n(dz;\hat{\theta}_n)$ consistently estimate $\alpha^{-1}\bar{K}(y)$. It follows that

$$\frac{n_A}{n}\int_{0<y\le x}\frac{\hat{H}_{n_A}(dy)}{\int_{z\ge y}[G(z;\hat{\theta}_n)]^{-1}\tilde{H}_n(dz;\hat{\theta}_n)}$$

is a consistent estimator of $A(x) = P(C_1^* \le x)$. Hence, the estimator $\tilde{F}_n(x;\hat{\theta}_n)$ is asymptotically equivalent to the solution of $U(x;\theta) = 0$, where

$$U(x;\theta) = \big[\tilde{F}_n(x;\theta)\big(1 - \frac{A(x)}{G(x;\theta)}\big)\big] - [\alpha^{-1}F(x)\bar{Q}(x)P(V^* \le x \le C_1^*)].$$

Since $P(C_1^* \ge V^*) = 1$, we have $1 - A(x)/G(x;\theta) = P(V^* \le x \le C_1^*)/G(x;\theta)$. It follows that $\tilde{F}_n(x;\hat{\theta}_n)$ is a consistent estimator of $\tilde{F}(x;\theta) = \alpha^{-1}G(x;\theta)F(x)\bar{Q}(x)$.

3. A Simulation Study

A simulation study is conducted to compare the performance of the semiparametric estimator $\hat{F}_n(x;\hat{\theta}_n)$ against that of the product-limit estimator $\hat{F}_n(x)$

3.1 Cases State

For all the cases considered, the T^*'s are exponential distribution: $F(x) = 1 - e^{-x}$ for $x > 0$, and the C_1^*'s are defined by $C_1^* = D^* + V^*$, where D^*'s are independent of V^*. We generate V^*, D^* and C_2^* from the following three cases:

Case 1 (Stationarity) :

The V^*'s are uniform distribution: $G(x;\theta) = x/\theta$ with varying parameters $\theta = 0.25, 1.0,$ and 4.0. The D^*'s are exponentially distributed: $Q^D(x) = 1 - e^{-x}$ for $x > 0$. The C_2^*'s are exponential distribution: $Q(x) = 1 - e^{-\beta_2 x}$ for $x > 0$, with varying parameters $\beta_2 = 1.0$, 2.0, and 4.0.

Case 2 :

The V^*'s are exponential distribution: $G(x;\theta) = 1 - e^{-\theta x}$ for $x > 0$, with varying parameters $\theta = 1.0, 4.0,$ and 8.0. The D^*'s are exponentially distributed: $Q^D(x) = 1 - e^{-\beta_d x}$ for $x > 0$, with varying parameters $\beta_d = 4.0, 8.0$. The C_2^*'s are exponential distribution: $Q(x) = 1 - e^{-2x}$ for $x > 0$.

Case 3 :

The V^*'s are exponential distribution: $G(x;\theta) = 1 - e^{-\theta x}$ for $x > 0$, with varying parameters $\theta = 1.0, 4.0$, and 8.0. The D^*'s are exponentially distributed: $Q^D(x) = 1 - e^{-\beta_d x}$ for $x > 0$, with varying parameters $\beta_d = 1.0, 4.0$. The C_2^*'s are exponential distribution: $Q(x) = 1 - e^{-0.25x}$ for $x > 0$.

For case 1, the θ is assumed to be known. For cases 2 and 3, the θ is assumed to be unknown. For all the cases, we consider the estimation of $F(0.5) = 0.39$, $F(1.0) = 0.63$ and $F(2.0) = 0.87$. The sample size is chosen as 200 and the replication is 3000 times. Tables 1 through 3 show the bias, standard deviation (std.) and squared root of the ratio of mean squared errors (denoted by eff) of the $\hat{F}_n(x;\hat{\theta}_n)$ to that of the product-limit estimator $\hat{F}_n(x)$. Tables 1 through 3 also show the proportion of truncation (α) and the proportion of type A and type B censoring ($p_A = n_A/n, p_B = n_B/n$). Based on the results of Tables 1 through 3, we have the following conclusions.

3.2 Simulation Results

Case 1: (see Table 1)

In terms of squared root of mean squared error ($\sqrt{mse}$), the semiparametric estimator $\hat{F}_n(x;\hat{\theta}_n)$ outperforms the product-limit estimator $\hat{F}_n(x)$ except in the case of light truncation and heavy censoring (i.e. $\alpha = 0.79$ $p_A = 0.26$ $p_B = 0.37$, eff=1.11). The ratio of the $\sqrt{mse}$ of $\hat{F}_n(x;\hat{\theta}_n)$ to that of $\hat{F}_n(x)$ varies between 0.27 to 1.11. For the estimation of $F(2.0)$, when truncation is light and censoring is heavy (e.g. $\alpha = 0.79$, $p_A = 0.52$, $p_B = 0.24$, and eff=0.27), the improvement of $\hat{F}_n(x;\hat{\theta}_n)$ can be very significant.

Case 2: (see Table 2)

For the estimation of $F(1.0)$ and $F(2.0)$, The estimator $\hat{F}_n(x;\hat{\theta}_n)$ outperforms the product-limit estimator for all the cases considered. The ratio of the $\sqrt{mse}$ of $\hat{F}_n(x;\hat{\theta}_n)$ to that of $\hat{F}_n(x)$ varies between 0.22 to 0.86. However, for the estimation of $F(0.5)$, the product-limit estimator can outperform $\hat{F}_n(x;\hat{\theta}_n)$. The ratio of the $\sqrt{mse}$ of $\hat{F}_n(x;\hat{\theta}_n)$ to that of $\hat{F}_n(x)$ varies between 0.85 to 1.11. For the estimation of $F(2.0)$, when truncation is light and censoring is heavy (e.g. $\alpha = 0.57$, $p_A = 0.42$, $p_B = 0.39$, and eff=0.22), the improvement of $\hat{F}_n(x;\hat{\theta}_n)$ can be very significant.

Case 3: (see Table 3)

The semiparametric estimator $\hat{F}_n(x;\hat{\theta}_n)$ outperforms the product-limit estimator for most of the case considered. The ratio of the $\sqrt{mse}$ of $\hat{F}_n(x;\hat{\theta}_n)$ to that of $\hat{F}_n(x)$ varies between 0.24 to 1.08. For the estimation of $F(2.0)$, when truncation is light and type A censoring is heavy (e.g. $\alpha = 0.87$, $p_A = 0.66$, $p_B = 0.06$, and eff=0.24), the improvement of $\hat{F}_n(x;\hat{\theta}_n)$ can be very significant.

Table 1. Simulation results for bias, std and $\sqrt{mse}$

of the estimators $\hat{F}_n(x;\hat{\theta})$ and $\hat{F}_n(x)$, Case 1

					$\hat{F}_n(0.5;\theta)$			$\hat{F}_n(0.5)$	
θ	β_2	α	p_A	p_B	bias	std	eff	bias	std
0.25	1.0	0.79	0.26	0.37	-0.000	0.051	0.87	-0.001	0.059
0.25	2.0	0.79	0.41	0.30	-0.001	0.055	0.88	-0.003	0.062
0.25	4.0	0.79	0.52	0.24	0.015	0.059	0.88	-0.002	0.070
1.00	1.0	0.43	0.14	0.42	-0.002	0.069	0.86	0.008	0.081
1.00	2.0	0.43	0.22	0.39	-0.001	0.073	0.84	0.006	0.087
1.00	4.0	0.43	0.29	0.35	-0.000	0.077	0.81	0.004	0.094
4.00	1.0	0.13	0.04	0.48	-0.001	0.072	0.97	-0.025	0.071
4.00	2.0	0.13	0.04	0.47	-0.010	0.074	0.89	-0.008	0.084
4.00	4.0	0.13	0.09	0.46	-0.010	0.077	0.87	-0.025	0.086

					$\hat{F}_n(1.0;\theta)$			$\hat{F}_n(1.0)$	
θ	β_2	α	p_A	p_B	bias	std	eff	bias	std
0.25	1.0	0.79	0.26	0.37	-0.001	0.057	0.81	0.003	0.070
0.25	2.0	0.79	0.41	0.30	0.006	0.079	0.96	-0.002	0.083
0.25	4.0	0.79	0.52	0.24	0.104	0.108	0.99	0.002	0.150
1.00	1.0	0.43	0.14	0.42	-0.001	0.055	0.90	0.006	0.062
1.00	2.0	0.43	0.22	0.39	0.002	0.061	0.85	0.003	0.071
1.00	4.0	0.43	0.29	0.35	-0.012	0.065	0.81	0.002	0.083
4.00	1.0	0.13	0.04	0.48	-0.009	0.057	0.97	-0.017	0.060
4.00	2.0	0.13	0.04	0.47	-0.007	0.057	0.89	-0.008	0.063
4.00	4.0	0.13	0.09	0.46	-0.011	0.064	0.77	-0.001	0.072

					$\hat{F}_n(2.0;\theta)$			$\hat{F}_n(2.0)$	
θ	β_2	α	p_A	p_B	bias	std	eff	bias	std
0.25	1.0	0.79	0.26	0.37	0.007	0.086	1.11	-0.007	0.077
0.25	2.0	0.79	0.41	0.30	0.017	0.078	0.59	-0.051	0.136
0.25	4.0	0.79	0.52	0.24	-0.084	0.067	0.27	-0.167	0.246
1.00	1.0	0.43	0.14	0.42	0.007	0.062	0.97	0.005	0.060
1.00	2.0	0.43	0.22	0.39	0.025	0.070	0.83	-0.010	0.090
1.00	4.0	0.43	0.29	0.35	0.032	0.051	0.45	-0.074	0.134
4.00	1.0	0.13	0.04	0.48	-0.005	0.041	0.92	-0.004	0.045
4.00	2.0	0.13	0.04	0.47	-0.009	0.047	0.81	-0.001	0.060
4.00	4.0	0.13	0.09	0.46	-0.011	0.059	0.73	-0.011	0.079

Table 2. Simulation results for bias, std and $\sqrt{mse}$

of the estimators $\hat{F}_n(x;\hat{\theta})$ and $\hat{F}_n(x)$, Case 2: $C_2^* \sim exp(2)$

					$\hat{F}_n(0.5;\hat{\theta}_n)$			$\hat{F}_n(0.5)$	
θ	β_d	α	p_A	p_B	bias	std	eff	bias	std
1.0	4.0	0.25	0.15	0.57	0.004	0.086	1.11	-0.006	0.077
1.0	8.0	0.25	0.18	0.55	0.031	0.090	1.06	-0.000	0.089
4.0	4.0	0.57	0.33	0.45	0.007	0.067	0.94	0.000	0.071
4.0	8.0	0.57	0.42	0.39	0.002	0.075	0.85	-0.002	0.088
8.0	4.0	0.73	0.42	0.39	0.017	0.074	1.00	-0.006	0.076
8.0	8.0	0.73	0.53	0.31	0.068	0.091	1.07	-0.002	0.105

					$\hat{F}_n(1.0;\hat{\theta}_n)$			$\hat{F}_n(1.0)$	
θ	β_d	α	p_A	p_B	bias	std	eff	bias	std
1.0	4.0	0.25	0.15	0.57	0.007	0.086	0.83	-0.001	0.104
1.0	8.0	0.25	0.18	0.55	0.039	0.066	0.49	0.006	0.154
4.0	4.0	0.57	0.33	0.45	0.052	0.114	0.74	-0.006	0.154
4.0	8.0	0.57	0.42	0.39	0.138	0.106	0.86	-0.044	0.196
8.0	4.0	0.73	0.42	0.39	0.103	0.120	0.67	-0.034	0.178
8.0	8.0	0.73	0.53	0.31	0.161	0.087	0.76	-0.113	0.212

					$\hat{F}_n(2.0;\hat{\theta}_n)$			$\hat{F}_n(2.0)$	
θ	β_d	α	p_A	p_B	bias	std	eff	bias	std
1.0	4.0	0.25	0.15	0.57	0.013	0.078	0.47	-0.078	0.149
1.0	8.0	0.25	0.18	0.55	0.042	0.043	0.27	-0.119	0.181
4.0	4.0	0.57	0.33	0.45	-0.039	0.093	0.38	-0.188	0.189
4.0	8.0	0.57	0.42	0.39	-0.036	0.065	0.22	-0.261	0.208
8.0	4.0	0.73	0.42	0.39	-0.061	0.093	0.35	-0.259	0.190
8.0	8.0	0.73	0.53	0.31	-0.064	0.079	0.25	-0.345	0.212

Table 3. Simulation results for bias, std and $\sqrt{mse}$

of the estimators $\hat{F}_n(x;\hat{\theta})$ and $\hat{F}_n(x)$, Case 3: $C_2^* \sim exp(0.25)$

					$\hat{F}_n(0.5;\hat{\theta}_n)$			$\hat{F}_n(0.5)$	
θ	β_d	α	p_A	p_B	bias	std	eff	bias	std
1.0	1.0	0.44	0.20	0.16	-0.006	0.072	0.83	-0.005	0.087
1.0	4.0	0.44	0.34	0.13	-0.003	0.087	0.98	-0.004	0.089
4.0	1.0	0.76	0.34	0.13	-0.002	0.052	0.85	-0.002	0.061
4.0	4.0	0.76	0.58	0.08	0.002	0.059	0.82	-0.001	0.072
8.0	1.0	0.87	0.39	0.12	0.002	0.045	1.07	-0.002	0.042
8.0	4.0	0.87	0.66	0.06	0.011	0.051	1.06	0.001	0.049

					$\hat{F}_n(1.0;\hat{\theta}_n)$			$\hat{F}_n(1.0)$	
θ	β_d	α	p_A	p_B	bias	std	eff	bias	std
1.0	1.0	0.44	0.20	0.16	-0.004	0.055	0.90	-0.001	0.061
1.0	4.0	0.44	0.34	0.13	-0.008	0.071	0.85	0.005	0.083
4.0	1.0	0.76	0.34	0.13	-0.004	0.049	0.92	-0.001	0.053
4.0	4.0	0.76	0.58	0.08	0.021	0.073	0.83	0.008	0.091
8.0	1.0	0.87	0.39	0.12	0.005	0.043	1.06	0.003	0.040
8.0	4.0	0.87	0.66	0.06	0.074	0.085	1.08	0.013	0.104

					$\hat{F}_n(2.0;\hat{\theta}_n)$			$\hat{F}_n(2.0)$	
θ	β_d	α	p_A	p_B	bias	std	eff	bias	std
1.0	1.0	0.44	0.20	0.16	0.001	0.038	0.88	-0.000	0.043
1.0	4.0	0.44	0.34	0.13	-0.016	0.054	0.56	-0.008	0.097
4.0	1.0	0.76	0.34	0.13	-0.002	0.051	0.96	0.002	0.053
4.0	4.0	0.76	0.58	0.08	0.038	0.047	0.39	-0.076	0.138
8.0	1.0	0.87	0.39	0.12	0.006	0.057	1.02	-0.028	0.052
8.0	4.0	0.87	0.66	0.06	0.030	0.048	0.24	-0.128	0.148

4. Concluding Remarks

The semiparametric estimate proposed in this article is designed to incorporate both information contained in the data and the available information on the truncation distribution, and are expected to have better performance than the nonparametric methods. Our simulation study indicates that under the semiparametric model $V^* \sim G(x;\theta)$, the semiparametric estimator $\hat{F}_n(x;\hat{\theta}_n)$ can perform much better than the product-limit estimator $\hat{F}_n(x)$. The truncation product-limit estimator, however, is still most appropriate under a totally nonparametric model. In practice, we can perform a formal goodness-of-fit test on the hypothesis $H_0 : V^* \sim G(x;\theta)$ using the method of Li and Doss (1993). Their method is based on a modified minimum chi-square estimator of θ, $\hat{\theta}_c$. For a fixed $\hat{\theta}_c$, an alternative semiparametric estimator, $\hat{F}_n(x;\hat{\theta}_c)$, can be obtained by maximizing $L(\tilde{F}, \tilde{Q}; \hat{\theta}_c)$. Further investigation is required for a comparison between $\hat{F}_n(x;\hat{\theta}_c)$ and $\hat{F}_n(x;\hat{\theta}_n)$.

References

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. Length-biased sampling with right censoring: an unconditional approach. J. Am. Statist. Ass. **2002**, 97, 201-209.

Asgharian, M. and Wolfson, D. B. Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. Ann. Statist., **2005**, 33, 2109-2131.

Li, G. and Doss, H. Generalized Person-Fisher Chi-square goodness-of-fit tests, with applications to models with life history data. Ann. Statist., **1993**, 21, 772-797.

Vardi, Y. Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. Biometrika, **1989**, 76, 751-761.

Vardi, Y. and Zhang, C.-H. Large sample study of empirical distributions in a random-multiplicative censoring model. Ann. Statist., **1992**, 20, 1022-1039.

Wang, M.-C. Product-limit estimates: a generalized maximum likelihood study. Communi. in Statist., Part A- Theory and Methods, **1987**, 6, 3117-3132.

Wang, M.-C. A semiparametric model for randomly truncated data. J. Am. Statist. Ass. **1989**, 84, 742-748.

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