# 東海大學統計學研究所

# 碩士論文

指導教授 : 沈葆聖 教授

THE NPMLE AS AN INVERSE-PROBABILITY-WEIGHTED AVERAGE



研究生 : 鄭志雄

中華民國 九十二 年 七 月

# THE NPMLE AS AN INVERSE-PROBABILITY-WEIGHTED AVERAGE

Chih-Hsiung Cheng

Dept. of Statistics

Tunghai University

Taichung, 40704

Taiwan, R. O. C.

June 10, 2003

# Contents

# THE NPMLE AS AN INVERSE-PROBABILITY-WEIGHTED AVERAGE

## ABSTRACT

For randomly censored data, Satten and Datta (2001) showed that the Kaplan-Meier estimator (known as a nonparametric maximum likelihood estimate (NPMLE)) can be expressed as an inverse-probability-weighted average. In this article, we consider the other two NPMLEs: the truncation NPMLE and the censoring-truncation NPMLE. For the data subject to left-truncation or both left-trucation and right-censoring, it is shown that these two NPMLEs can be expressed as inverse-probability-weighted averages.

Keywords: NPMLE; Inverse-probability-weighted average.

# Chapter 1. INTRODUCTION

The Kaplan-Meier estimator (product-limit estimator (PLE)) for the survival function of randomly censored time-to-event data (Kaplan and Meier (1)) is often introduced as the maximizer of a nonparametric maximum likelihood (see Kalbfleisch and Prentice (2); Wang (3) ). In a series of papers, Robins and coworkers proposed a class of estimators using a data-reweighting scheme (Robins and Rotnitzky (4); Robins (5); Robins and Finkelstein (6)). An outcome of their approach applied to survival analysis is an inverse-probability-of-censoring representation of the Kaplan-Meier estimator. Satten and Datta (7) give two demonstrations of this representation. In this article, we consider the other two PLEs: the truncation PLE and the censoring-truncation PLE. These two PLEs were introduced by Lynden-Bell (8) and Cox and Oakes (9), respectively, and their large sample properties were studied by Woodroofe (10), Wang et al. (11) and Tsai et al. (12). In Section 2 and 3, it will be shown that for the data subject to left-truncation or both left-trucation and right-censoring, these two PLEs can be expressed as inverse-probability-weighted averages.

## Chapter 2. LEFT-TRUNCATED DATA

Let $U^*$ and $V^*$ be the target and truncation variables with distribution functions $F$ and $G$ respectively. Assume that $U^*$ and $V^*$ are independent. For left-truncated data, both $U^*$ and $V^*$ are observable only when $U^* \geq V^*$. Let $(U_1, V_1), \ldots, (U_n, V_n)$ denote the truncated sample. Hence, $H(u, v) = P(U_i \leq u, V_i \leq v) = P(U^* \leq u, V^* \leq v | U^* \geq V^*)$. Let $I_{[A]}$ be the indicator function of the event $A$. Let $N_F(u) = \sum_{i=1}^n I_{[U_i \leq u]}$, $N_G(v) = \sum_{i=1}^n I_{[V_i \leq v]}$, and $R_n(u) = N_G(u) - N_F(u-) = \sum_{i=1}^n I_{[V_i \leq u \leq U_i]}$. The truncation PLEs of $F$ and $G$ can be viewed as a nonparametric method for dealing with delayed entry of uncensored life table data, as well as truncated astronomy data (see Woodroofe (10); Wang et al. (11); He and Yang (13)). The follwing examples describes situtations where the models of left truncation are appropriate.

**Example 1 (retirement data)**:

Channing House is a retirement center located in Palo Alto, California. Data on ages at death of 462 individuals (97 males and 365 females), who were in residence during the period January 1964 to July 1975, has been reported by Hyde (1980). The life lengths in this data set are left-truncated because an individual must survive to a sufficient age to enter the retirement community. The truncation variable $V^*$, is then the potential patient's age at entry, and the target variable $U^*$, is the patient's age at death. Obviously we can only observe $(U^*, V^*)$ if $U^* \geq V^*$.

**Example 2 (AIDS blood-transfusion data)**:

The blood transfusion related AIDS data given by Kalbfleisch and Lawless (1989). They gives infection times $V^*$, in months with 1 representing January 1978, incubation times $T$ in months, and age in years for 34 'children' aged 0 to 4 years, 120 'adults' aged 5 to 59 years, and 141 'elderly' aged 60 and over, who were infected by contaminated

blood transfusions and developed AIDS by 1 July 1986. Let $U^* = 102 - T$. The truncation effect comes from the fact that we only observed over the period $(0, 102]$. An individual is observed if and only if $T + V^* \leq 102$ or $V^* \leq U^*$.

Let $U_{(1)} < U_{(2)} < \cdots < U_{(r)}$ denote the distinct ordered statistics of the sample $U_i's$. Let $d_i = N_F(U_{(i)}) - N_F(U_{(i)}-)$ denote the number of failure times at $U_{(i)}$ for $i = 1, \ldots, r$. Similarly, let $V_{(1)} < V_{(2)} < \cdots < V_{(q)}$ be the distinct order statistics of sample $V_1, V_2, \ldots, V_n$, and $e_j = N_G(V_{(j)}) - N_G(V_{(j)}-)$ denote the number of truncation times at $V_{(j)}$. A necessary and sufficient condition for the existence of the nonparametric maximum likelihood estimate (NPMLE) of $F(x)$ is $R_n(U_{(i)}) > d_i$ for $i = 1, \ldots, r$, for the existence of the NPMLE of $G(x)$ is $R_n(V_{(j)}) > e_j$ for $j = 1, \ldots, q-1$ (see Wang et al. (11)). Under these regularity conditions, the NPMLEs of $F(x)$ and $G(x)$ are uniquely determined and given by

$$\hat{F}_n(x) = 1 - \prod_{u \leq x} \Big[1 - \frac{dN_F(u)}{R_n(u)}\Big],$$

and

$$\hat{G}_n(x) = \prod_{v > x} \Big[1 - \frac{dN_G(v)}{R_n(v)}\Big],$$

where $dN_F(u) = N_F(u) - N_F(u-)$ and $dN_G(v) = N_G(v) - N_G(v-)$.

Under the semiparametric model, $V^*$ is assumed to have distribution function $G(y; \theta)$, where $G$ is specified, $\theta \in \Theta$ and $\theta$ can be a vector. For the semiparametric model, the MLE of $F(x)$, derived by Wang (14), is

$$\big(\sum_i 1/G(U_i; \hat{\theta})\big)^{-1} \sum_i \frac{I_{[U_i \leq x]}}{G(U_i; \hat{\theta})} \tag{2.1}$$

where $\hat{\theta}$ is the MLE of the conditional likelihood function $U_i$'s given $V_i$'s. Note that when $\theta$ is known, the weighted average (2.1) is actually the MLE described by Vardi (15), with $G$ a weight function. Please refer to Vardi (15) for selection-bias models with known weights.

We shall give two demonstrations of the equivalence of the inverse-probability-of-truncation weighted estimate and the Lynden-Bell's (8) estimator. The first, substitution of $\hat{G}_n(U_i)$ for $G(U_i;\theta)$ in (2.1) leads to an inverse-probability-of-truncation weighted estimator

$$\hat{F}_w(x) = \big(\sum_i 1/\hat{G}_n(U_i)\big)^{-1} \sum_i \frac{I_{[U_i \le x]}}{\hat{G}_n(U_i)}.$$

The following theorem shows the equivalence of $\hat{F}_w$ and $\hat{F}_n$.

**Theorem 2.1.** $\hat{F}_w = \hat{F}_n$

**Proof:**

Note that both $\hat{F}_w$ and $\hat{F}_n$ are step right-continuous functions. Thus, $\hat{F}_w$ and $\hat{F}_n$ are the same if the magnitudes of the jumps in the two functions are equal. The jump in $\hat{F}_w$ at time $U_{(i)}$ is given by

$$\hat{F}_w(U_{(i)}) - \hat{F}_w(U_{(i-1)}) = \frac{d_i/\hat{G}_n(U_{(i)})}{\sum_{j=1}^r d_j/\hat{G}_n(U_{(j)})}.$$

Now, by Corollary 2.4 of He and Yang (13), we have

$$\frac{d_i/\hat{G}_n(U_{(i)})}{\sum_{j=1}^r d_j/\hat{G}_n(U_{(j)})} = \frac{d_i[1-\hat{F}_n(U_{(i-1)})]/R_n(U_{(i)})}{\sum_{j=1}^r d_j[1-\hat{F}_n(U_{(j-1)})]/R_n(U_{(j)})}.$$

Since

$$\sum_{j=1}^r \frac{d_j[1-\hat{F}_n(U_{(j-1)})]}{R_n(U_{(j)})} = \sum_{j=1}^r \prod_{k=1}^{j-1} \Big(\frac{R_n(U_{(k)})-d_k}{R_n(U_{(k)})}\Big)\Big(\frac{d_j}{R_n(U_{(j)})}\Big)$$

$$= \sum_{j=1}^r \prod_{k=1}^{j-1} \Big(\frac{R_n(U_{(k)})-d_k}{R_n(U_{(k)})}\Big)\Big[1-\frac{R_n(U_{(j)})-d_j}{R_n(U_{(j)})}\Big] = \sum_{j=1}^r [\hat{F}_n(U_{(j)}) - \hat{F}_n(U_{(j-1)})] = 1,$$

we have

$$\hat{F}_w(U_{(i)}) - \hat{F}_w(U_{(i-1)}) = \frac{d_i[1-\hat{F}_n(U_{(i-1)})]}{R_n(U_{(i)})} = \hat{F}_n(U_{(i)}) - \hat{F}_n(U_{(i-1)}).$$

Thus, $\hat{F}_w$ and $\hat{F}_n$ are the same.

Next, we introduce an inverse-probability-of-truncation weighted estimator of $F(x)$ that makes no reference to $\hat{G}_n$. We then show that this estimator is identical to the Lynden-Bell's (8) product-limit estimate. We simultaneously estimate $F(x)$ and $G(y)$ using coupled inverse-prbability-of-truncation weighted estimators. Let $\hat{F}_c(x)$ and $\hat{G}_c(x)$ be given by

$$\hat{F}_c(x) = \big(\sum_{i=1}^{n} 1/\hat{G}_c(U_i)\big)^{-1} \sum_{i=1}^{n} \frac{I_{[U_i \le x]}}{\hat{G}_c(U_i)},$$

and

$$\hat{G}_c(x) = \big(\sum_{i=1}^{n} 1/[1-\hat{F}_c(V_i-)]\big)^{-1} \sum_{i=1}^{n} \frac{I_{[V_i \le x]}}{[1-\hat{F}_c(V_i-)]}.$$

We shall show that $\hat{F}_c$ and $\hat{F}_n$ are equivalent.

**Theorem 2.2.** $\hat{F}_c = \hat{F}_n$

**Proof:**

Note that

$$\hat{F}_c(U_{(i)}) = \big(\sum_{s=1}^{r} d_s/\hat{G}_c(U_{(s)})\big)^{-1} \sum_{s=1}^{i} \frac{d_s}{\hat{G}_c(U_{(s)})}, \tag{2.2}$$

and

$$\hat{G}_c(V_{(j)}) = \big(\sum_{t=1}^{q} e_t/[1-\hat{F}_c(V_{(t)}-)]\big)^{-1} \sum_{t=1}^{j} \frac{e_t}{[1-\hat{F}_c(V_{(t)}-)]}. \tag{2.3}$$

Denote the jump in $\hat{F}_c(U_{(i)})$ and $\hat{G}_c(V_{(j)})$ by $f_i$ and $g_j$, respectively. By (2.2) and (2.3), we have

$$f_i = \frac{d_i/\sum_{V_{(i')} \le U_{(i)}} g_{i'}}{\sum_{s=1}^{r} d_s/\sum_{V_{(s')} \le U_{(s)}} g_{s'}}, \tag{2.4}$$

and

$$g_j = \frac{e_j/\sum_{U_{(j')} \ge V_{(j)}} f_{j'}}{\sum_{t=1}^{q} e_t/\sum_{U_{(t')} \ge V_{(t)}} f_{t'}}. \tag{2.5}$$

By (2.4) and (2.5), we have

$$f_i = \frac{d_i}{\sum_{V_{(i')} \le U_{(i)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}}} \left[ \sum_{s=1}^{r} \frac{d_s}{\sum_{V_{(i')} \le U_{(s)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}}} \right]^{-1}.$$

The masses in the $\hat{F}_n$ are the maximizers of the following likelihood (see Wang (3)):

$$L = \prod_{i=1}^{n} \frac{f(U_i)}{\sum_{U_{i''} \ge V_i} f(U_{i''})} = \prod_{s=1}^{r} f_s^{d_s} \prod_{t=1}^{q} \left[ \sum_{U_{(i'')} \ge V_{(t)}} f_{i''} \right]^{-e_t}.$$

Following Turnbull (16), note that $\{f_i, 1 \le i \le r\}$ solves this maximization problem if

$$D_i = \frac{\partial lnL}{\partial f_i} - \sum_{s=1}^{r} f_s \frac{\partial lnL}{\partial f_s} = 0,$$

and $\sum_{i=1}^{r} f_i = 1$.

First,

$$\sum_{s=1}^{r} f_s \frac{\partial lnL}{\partial f_s} = \sum_{i=1}^{r} d_i - \sum_{s=1}^{r} f_s \sum_{V_{(i')} \le U_{(s)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}}$$

$$= \sum_{i=1}^{r} d_i - \sum_{t=1}^{q} e_t \sum_{U_{(i')} \ge V_{(t)}} \frac{f_{i'}}{\sum_{U_{(i'')} \ge V_{(t)}} f_{i''}} = \sum_{i=1}^{r} d_i - \sum_{t=1}^{q} e_t = 0.$$

Next,

$$\frac{\partial lnL}{\partial f_i} = \sum_{i=1}^{r} \frac{f_i}{d_i} - \sum_{V_{(i')} \le U_{(i)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}}.$$

Hence, for $i = 1, \ldots, r$, $\partial lnL / \partial f_i = 0$ implies that

$$f_i = \frac{d_i}{\sum_{V_{(i')} \le U_{(i)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}}}. \tag{2.6}$$

Notice that (2.6) stays the same if $f_i$ $(i = 1, \ldots, r)$ is multiplied by any positive constant. The constraint $\sum_{s=1}^{r} f_s = 1$ establishes the equivalence of $\hat{F}_c$ and $\hat{F}_n$

The proof of the equivalence of $\hat{G}_c$ and $\hat{G}_n$ is similar to that of $\hat{F}_c$ and $\hat{F}_n$, and is omitted. In certain cases, there are possibilities of a right truncation and the relevent

cases are well described by Wang (14). The approach for right-truncated data is easily dealt by reversing the roles of target and truncation variables.

## Chapter 3. LEFT-TRUNCATED AND RIGHT-CENSORED DATA

Let $(U_i^*, C_i, V_i^*)$ be i.i.d. random vectors such that $(C_i, V_i^*)$ is independent of $U_i^*$. It will be assumed throughout this section that $C_i \geq V_i^*$. Let $F$, $Q$ and $G$ denote the common distribution function of $U_i^*$, $C_i$ and $V_i^*$, respectively. For left-truncated and right-censored data, one can observe nothing if $U_i^* < V_i^*$ and observe $(X_i^*, \delta_i^*)$, with $X_i^* = \min(U_i^*, C_i)$ and $\delta_i^* = I_{[U_i^* \leq C_i]}$, if $U_i^* \geq V_i^*$. Data of this kind often arise in epidemiology and individual follow-up study (see Wang (17)). An example of left-truncated and right-censored data is given as follows.

**Example:**

Let the initiating event correspond to the onset of a certain disease, and let the terminating event correspond to the death of an individual. Suppose that the disease population in a certain city is a representative sample from a large disease population. The target interest of a research project is to study the natural history of the disease for individuals who developed the disease during the calender time period $(\tau_1, \tau_2)$, $\tau_1 < \tau_2$. Consider the sampling under which all of the individuals in the city who have developed the disease between $\tau_1$ and $\tau_2$ and survived past the calender time $\tau$ $(\tau_2 < \tau)$ are recruited at the time $\tau$ for a prospective follow-up study. Suppose that the researchers are able to identify the calendar times of disease onset and death for those in the city who have developed the disease between $\tau_1$ and $\tau_2$ and died before $\tau$ from, for example, information provided by health department. Let $U_i^*$ denote the time from onset of disease to death, $V_i^*$ corrspond to the time from the onset of disease to $\tau$, and $C_i^*$ correspond to the time from the onset of disease to censoring. Clearly, the calender time of the potential censoring point must be greater than $\tau$, since only those individuals in the follow-up study might be observed subject to right censoring. Therefore, $P(V_i^* < C_i^*) = 1$, and we observe $(X_i^*, \delta_i^*)$ if $U_i^* \geq V_i^*$.

**Notation**

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

Let $U_{(1)} < U_{(2)} < \cdots < U_{(r)}$ be the distinct ordered failure times and $d_s$ be the number of failure times at $U_{(s)}$ for $s = 1, \ldots, r$.

Similarly, let $V_{(1)} < V_{(2)} < \cdots < V_{(q)}$ be the distinct ordered truncation times and $e_t$ be the number of truncation times at $V_{(t)}$ for $t = 1, \ldots, q$.

Let $C_{(1)} < C_{(2)} < \cdots < C_{(h)}$ be the distinct ordered censoring times and $c_l$ be the number of censoring times at $C_{(l)}$ for $l = 1, \ldots, h$.

For each $V_{(t)}$ $(t = 1, \ldots, q)$, let $C_{(1(t))} < C_{(2(t))} < \cdots < C_{(h(t))}$ be the distinct ordered censoring times and $c_{l(t)}$ be the number of censoring times at $C_{(l(t))}$ for $l = 1, \ldots, h(t)$.

For each $V_{(t)}$ $(t = 1, \ldots, q)$, let $U_{(1(t))} < U_{(2(t))} < \cdots < U_{(r(t))}$ be the distinct ordered failure times and $d_{s(t)}$ be the number of censoring times at $U_{(s(t))}$ for $s = 1, \ldots, r(t)$.

Let $Q(x|v) = P(C_i \leq x | V_i^* = v)$ denote the conditional distribution function of $C_i$ given $V_i^* = v$.

Let $dF(x) = F(x) - F(x-)$, $dG(x) = G(x) - G(x-)$, $dQ(x|v) = Q(x|v) - Q(x - |v)$ and $p = P(U_i^* \geq V_i^*)$.

The likelihood function $L$ can be decompose into three factors (see Wang (17), Gross and Lai (18)), yielding

$$L = \prod_{i=1}^{n} \Big\{ dF(X_i) dG(V_i) [1 - Q(X_i - |V_i)/p \Big\}^{\delta_i} \times \prod_{i=1}^{n} \Big\{ dQ(X_i|V_i) dG(V_i) [1 - F(X_i)]/p \Big\}^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left\{ \frac{[dF(X_i)]^{\delta_i} [1 - F(X_i)]^{1-\delta_i}}{1 - F(V_i-)} \right\} \times \left\{ \prod_{t=1}^{q} \left[ \frac{dG(V_{(t)})[1 - F(V_{(t)}-)]}{p} \right]^{e_t} \right\}$$

$$\times\left\{\prod_{t=1}^{q}\left[\prod_{V_i=V_{(t)}}[1-Q(X_i-|V_{(t)})]^{\delta_i}[dQ(X_i|V_{(t)})]^{1-\delta_i}\right]\right\}=L_1L_2L_3,$$

where $L_1$, $L_2$, and $L_3$ represent the likelihoods in the first, second, and third brace, respectively. Note that $L_1$, $L_2$, and $L_3$ can be written as

$$L_1=\prod_{s=1}^{r}f_s^{d_s}\prod_{l=1}^{h}\left[\sum_{U_{(i'')}\geq C_{(l)}}f_{i''}\right]^{c_l}\prod_{t=1}^{q}\left[\sum_{U_{(i')}\geq V_{(t)}}f_{i'}\right]^{-e_t},$$

$$L_2=\prod_{t=1}^{q}\left[g_t\sum_{U_{(i')}\geq V_{(t)}}f_{i'}\right]^{e_t}\left[\sum_{t=1}^{q}g_t\sum_{U_{(i')}\geq V_{(t)}}f_{i'}\right]^{-n},$$

and

$$L_3=\prod_{t=1}^{q}\left\{\prod_{l(t)=1}^{h(t)}q_{l(t)}^{c_{l(t)}}\prod_{s(t)=1}^{r(t)}\left[\sum_{C_{(l'(t))}\geq U_{(s(t))}}q_{l'(t)}\right]^{d_{s(t)}}\right\}^{e_t},$$

where $f_s$ $(s=1,\ldots,r)$ and $g_t$ $(t=1,\ldots,q)$ denote the jump in $\tilde{F}_n(U_{(s)})$ and $\tilde{G}_n(V_{(t)})$ respectively, and $q_{l(t)}$ $(l=1,\ldots,h;t=1\ldots,q)$ denote the jump in $\tilde{Q}(C_{(l)}|V_{(t)})$.

Let $\tilde{R}_n(u)=\sum_{i=1}^{n}I_{[V_i\leq u\leq X_i]}$ and $\tilde{N}_F(u)=\sum_{i=1}^{n}I_{[X_i\leq u,\delta_i=1]}$. A necessary and sufficient condition for the existence of the NPMLE of $L_1$ is $\tilde{R}_n(U_{(s)})>d_s=\tilde{N}_F(U_{(s)})-\tilde{N}_F(U_{(s)}-)$ for $s=1,\ldots,r$. Under this regularity condition, the NPMLE of $F(x)$ from $L_1$ is uniquely determined and given by

$$\tilde{F}_n(x)=1-\prod_{u\leq x}\Big[1-\frac{d\tilde{N}_F(u)}{\tilde{R}_n(u)}\Big],$$

where $d\tilde{N}_F(u)=\tilde{N}_F(u)-\tilde{N}_F(u-)$.

Wang (3) showed that $\tilde{F}_n(x)$ is also the unique NPMLE of the full likelihood $L$. Based on $L_2$, the NPMLE of $G(x)$ is uniquely determined and given by

$$\tilde{G}_n(y)=\left[\sum_{t=1}^{q}\frac{e_t}{1-\tilde{F}_n(V_{(t)}-)}\right]^{-1}\sum_{t=1}^{q}\frac{e_tI_{[V_{(t)}\leq y]}}{1-\tilde{F}_n(V_{(t)}-)}.$$

Next, let $\tilde{R}_n^t(u)=\sum_{i=1}^{n}I_{[V_i\leq u\leq X_i,V_i=V_{(t)}]}$ and $\tilde{N}_Q^t(u)=\sum_{i=1}^{n}I_{[X_i\leq u,\delta_i=0V_i=V_{(t)}]}$. For each $V_{(t)}$, a necessary and sufficient condition for the existence of the NPMLE of $Q(x|V_{(t)})$

is $\tilde{R}_n^t(C_{(l(t))}) > c_{l(t)} = \tilde{N}_Q^t(C_{l(t)}) - \tilde{N}_Q^t(C_{l(t)}-)$ for $l = 1, \ldots, h(t)$. Under this regularity condition, the NPMLE of $Q(x|V_{(t)})$ from $L_3$ is uniquely determined and given by

$$\tilde{Q}_n(x|V_{(t)}) = 1 - \prod_{u \le x} \Big[1 - \frac{d\tilde{N}_Q^t(u)}{\tilde{R}_n^t(u)}\Big]$$

where $d\tilde{N}_Q^t(u) = N_Q^t(u) - N_Q^t(u-)$.

When $\tilde{Q}_n(x|V_{(t)})$ exists for all $V_{(t)}$'s, the nonparametric MLE of $Q$ (denoted by $\tilde{Q}_n$) can be written as

$$\tilde{Q}_n(x) = \sum_{t=1}^{q} \tilde{Q}_n(x|V_{(t)})[\tilde{G}_n(V_{(t)}) - \tilde{G}_n(V_{(t-1)})].$$

Next, we will show that $\tilde{F}_n$ can also be expressed as inverse-probability-weighted average. Due to the presence of censoring, the first procedure used in Theorem 2.1 is not feasible for left-truncated and right-censored data. However, we can consider the inverse-probability-weighted estimators by simultaneously estimating $F$, $G$ and $Q$. Let $\hat{F}_e(x)$, $\hat{G}_e(x)$ and $\hat{Q}_e(x)$ be given by

$$\hat{F}_e(x) = \left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{\delta_i I_{[X_i \le x]}}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)},$$

$$\hat{G}_e(x) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{I_{[V_i \le x]}}{1 - \hat{F}_e(V_i-)},$$

and

$$\hat{Q}_e(x) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{(1-\delta_i) I_{[X_i \le x]}}{1 - \hat{F}_e(X_i-)}.$$

Thus, we have

$$\hat{F}_e(U_{(i)}) = \left[\sum_{s=1}^{r} \frac{d_s}{\hat{G}_e(U_{(s)}) - \hat{Q}_e(U_{(s)}-)}\right]^{-1} \sum_{s=1}^{i} \frac{d_s}{\hat{G}_e(U_{(s)}) - \hat{Q}_e(U_{(s)}-)}, \tag{3.1}$$

$$\hat{G}_e(V_{(j)}) = \left[\sum_{t=1}^{q} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}\right]^{-1} \sum_{t=1}^{j} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}, \tag{3.2}$$

$$\hat{Q}_e(C_{(k)}) = \left[\sum_{t=1}^{q} \frac{e_t}{1-\hat{F}_e(V_{(t)}-)}\right]^{-1} \sum_{l=1}^{k} \frac{c_l}{1-\hat{F}_e(C_{(l)}-)}. \tag{3.3}$$

Similar to Theorem 2.2, the following theorem shows the equivalence of $\tilde{F}_n$ and $\hat{F}_e$.

**Theorem 3.1.** $\hat{F}_e = \tilde{F}_n$

**Proof:**

Note that the masses in the estimator $\tilde{F}_n$ are the maximizers of the the likelihood $L_1$ subject to $\sum_{s=1}^{r} f_s = 1$. Hence,

$$\frac{\partial lnL_1}{\partial f_i} - \sum_{s=1}^{r} f_s \frac{\partial lnL_1}{\partial f_s} = 0.$$

First,

$$\sum_{s=1}^{r} f_s \frac{\partial lnL}{\partial f_s} = \sum_{i=1}^{r} d_i - \sum_{s=1}^{r} f_s \sum_{V_{(i')} \le U_{(s)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}} + \sum_{s=1}^{r} f_s \sum_{C_{(l')} \le U_{(s)}} \frac{g_{l'}}{\sum_{U_{(i'')} \ge C_{(l')}} f_{i''}}$$

$$= \sum_{i=1}^{r} d_i - \sum_{t=1}^{q} e_t \sum_{U_{(i')} \ge V_{(t)}} \frac{f_{i'}}{\sum_{U_{(i'')} \ge V_{(t)}} f_{i''}} + \sum_{l=1}^{h} q_l \sum_{U_{(i')} \ge C_{(l)}} \frac{f_{i'}}{\sum_{U_{(i'')} \ge C_{(l)}} f_{i''}}$$

$$= \sum_{i=1}^{r} d_i - \sum_{t=1}^{q} e_t + \sum_{l=1}^{h} q_l = 0.$$

Next,

$$\frac{\partial lnL_1}{\partial f_i} = \sum_{i=1}^{r} \frac{f_i}{d_i} - \sum_{V_{(i')} \le U_{(i)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}} + \sum_{C_{(l')} \le U_{(i)}} \frac{g_{l'}}{\sum_{U_{(i'')} \ge C_{(l')}} f_{i''}}.$$

Hence, for $i = 1, \ldots, r$, $\partial lnL_1/\partial f_i = 0$ implies that

$$f_i = \frac{d_i}{\sum_{V_{(i')} \le U_{(i)}} \frac{e_{i'}}{\sum_{U_{(i'')} \ge V_{(i')}} f_{i''}} - \sum_{C_{(l')} \le U_{(i)}} \frac{q_{l'}}{\sum_{U_{(i'')} \ge C_{(l')}} f_{i''}}}.$$

By (3.1), (3.2), (3.3) and the constraint $\sum_{s=1}^{r} f_s = 1$, it follows that $f_i$ is also the jump in $\hat{F}_e(U_{(i)})$, and $\hat{F}_e$ and $\tilde{F}_n$ are equivalent.

By Theorem 3.1 and (3.2), it follows that $\hat{G}_e$ and $\tilde{G}_n$ are equivalent.

Now, the question left is 'Does the equivalence of $\hat{Q}_e$ and $\tilde{Q}_n$ hold ?'. First, for each $t$, let $h(t)$ denote the number of the distinct censoring times. Note that $\{q_{l(t)}, 1 \leq l \leq h(t)\}$ solves this maximization problem if

$$D_{lt} = \frac{\partial lnL_3}{\partial q_{l(t)}} - \sum_{l=1}^{h(t)} q_{l(t)} \frac{\partial lnL_3}{\partial q_{l(t)}} = 0,$$

and $\sum_{l=1}^{h(t)} q_{l(t)} = 1$. Some algebra shows that

$$q_{l(t)} = \frac{c_{l(t)}}{e_t - \sum_{\substack{U_{(l')} \leq C_{(l)} \\ d_{l'(t)} > 0}} \frac{d_{l'(t)}}{\sum_{\substack{C_{(l'')} \geq U_{(l')} \\ c_{l''(t)} > 0}} q_{l''(t)}}}.$$

Hence, the jump in $\tilde{Q}_n(C_{(l)})$ (denoted by $\tilde{q}_l$) can be written as

$$\tilde{q}_l = \left[\sum_{t=1}^{q} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}\right]^{-1} \sum_{t=1}^{q} q_{l(t)} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}$$

$$= \left[\sum_{t=1}^{q} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}\right]^{-1} \sum_{t=1}^{q} \frac{c_{l(t)}}{\tilde{S}_n(C_{(l)}|V_{(t)}-)[1 - \hat{F}_e(V_{(t)}-)]}, \tag{3.4}$$

where

$$\tilde{S}_n(C_{(l)}|V_{(t)}-) = 1 - e_t^{-1} \sum_{\substack{U_{(l')} \leq C_{(l)} \\ d_{l'}(t) > 0}} \frac{d_{l'(t)}}{\sum_{\substack{C_{(l'')} \geq U_{(l')} \\ c_{l''(t)} > 0}} q_{l''(t)}}.$$

By Satten and Datta (7), $\tilde{S}_n(C_{(l)}|V_{(t)}-)$ is an inverse-probability-of-censoring weighted estimator of the conditional survival function $P(U_i^* > C_{(l)}|U_i^* \geq V_{(t)})$. Hence,

$\tilde{S}_n(C_{(l)}|V_{(t)}-)[1 - \hat{F}_e(V_{(t)}-)]$ actually estimates $1 - F(C_{(l)})$.

Now, the jump in $\hat{Q}_e(C_{(l)})$ (denoted by $\hat{q}_l$) can be written as

$$\hat{q}_l = \left[\sum_{t=1}^{q} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}\right]^{-1} \sum_{t=1}^{q} \frac{c_{l(t)}}{1 - \hat{F}_e(C_{(l)})}. \tag{3.5}$$

By the expression of (3.4) and (3.5), we see that the equivalence of $\hat{Q}_e$ and $\tilde{Q}_n$ does not hold. When the bivariate distribution of $(C_i, V_i^*)$ is continuous, no more than one censored $C_{(l)}$ can be associated with each $V_{(t)}^*$ and therefore the NPMLE of $Q(z|V_{(t)})$ does not exist (i.e. $\tilde{R}_n^t(C_{(l)}) - c_{l(t)} = 0$). To circumvent this difficulty, as discussed by Gross and Lai (18), one way is to assume independence between $V_i^*$ and $C_i - V_i^*$. Under this assumption, the likelihood $L_3$ is reduced to (see Wang (17))

$$\prod_{i=1}^{n} [1 - W(X_i)]^{\delta_i} [dW(X_i)]^{1-\delta_i},$$

where $W(x)$ denotes the distribution function of $C_i - V_i^*$ and $dW(x) = W(x) - W(x-)$. It follows that the NPMLE of $W(x)$ is (see Wang (17), Gross and Lai (18))

$$\tilde{W}_n(x) = 1 - \prod_{u \le x} \Big[1 - \frac{d\tilde{N}_w(u)}{\tilde{R}_w(u)}\Big],$$

where $d\tilde{N}_w(u) = \sum_{i=1}^{n} I_{[X_i - V_i = u, \delta_i = 0]}$ and $\tilde{R}_w(u) = \sum_{i=1}^{n} I_{[X_i - V_i \ge u]}$.

Another apporach is to impose suitable smoothness assumptions (see Gross and Lai (18)) on the bivariate distribution of $V_i^*$ and $C_i$ so that $Q(x|v)$ is well approximated by $P(C_i \le x | G(v) - \Delta_n \le G(V_i^*) \le G(v) + \Delta_n)$ which can be consistenly estimated when $\Delta_n$ approaches 0 at a certain rate depending on the sample size $n$ as $n \to \infty$.

# Chapter 4. DISCUSSION

Following recent work by Satten and Datta (7), this article extends the weighted-average form of the PLEs to the data subject to left-truncation or both left-trucation and right-censoring. In survival analysis, the weighted-average approach can lead to useful generalizations, primarily to more general censoring or truncated models where censoring or truncation need not be identically distributed. Also, as seen in literature of censoring models, one application of the inverse-probability-weighting is to handle informative censoring in collection of survival data (see Robins (5), Robins and Finkelstein (6), Satten, Datta and Robins (19)). The weighted-average approach can be extended to the situations when the independent censoring or truncation assumption is violated. For example, assume for the $i^{th}$ person, data is available on time dependent covariates $Z_{ij}(x)$, $1 \leq j \leq J$ that may affect both failure and censoring times. Let $\bar{Z}_i(x)$ denote the information on all values of $Z_{ij}(x)$ between 0 and $x$. Assume that the $i^{th}$ person's hazard of being censored at time $x$ does not depend on $U_i^*$ given $\bar{Z}_i(x)$ and $X_i^* \geq x$. Let

$$Q_i(x) = \prod_{s \leq x} \Big[ 1 - d\Lambda_c[s|\bar{Z}_i(s)] \Big],$$

where $\Lambda_c[s|\bar{Z}_i(s)]$ denotes the cumlative hazard function of censoring times. For right-censored data, $\Lambda_c[s|\bar{Z}_i(s)]$ can be estimated using proportional hazard model or Aalen's additive hazard model. Based on Aalen's additive model, Satten, Datta and Robins (19) proposed a product-limit type estimator using a data-reweighting scheme. However, for left-truncated and right-censored data, further investigation is needed in estimating $\Lambda_c[s|\bar{Z}_i(s)]$. Suppose that there exists a reasonable estimator $\hat{\Lambda}_c[s|\bar{Z}_i(s)]$ for left-truncated and right-censored data, an inverse-probability-weighted estimator can be given by

$$\hat{F}_d(x) = \left[ \sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_i(X_i-)} \right]^{-1} \sum_{i=1}^{n} \frac{\delta_i I_{[X_i \leq x]}}{\hat{G}_e(X_i) - \hat{Q}_i(X_i-)}, \tag{4.1}$$

where

$$\hat{Q}_i(x) = 1 - \prod_{s \le x} \Big[ 1 - d\hat{\Lambda}_c[s|\bar{Z}_i(s)] \Big].$$

For the right-censored data, (4.1) is reduced to

$$\hat{F}_d(x) = \left[ \sum_{i=1}^{n} \frac{\delta_i}{1 - \hat{Q}_i(X_i-)} \right]^{-1} \sum_{i=1}^{n} \frac{\delta_i I_{[X_i \le x]}}{1 - \hat{Q}_i(X_i-)}. \tag{4.2}$$

It can be shown that (4.2) is different from the product-limit type estimator proposed by Satten, Datta and Robins (19). The comparison between these two estimators requires further investigation.

## REFERENCES

(1) Kaplan, E. L.; Meier, P. Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc., **1958**, *53*, 457-481.

(2) Kalbfleisch, J.; Prentice, R. The statistical analysis of failure time data, New York: Wiley, **1980**.

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

(4) Robins, J. M. and Rotnitzky, A. Rocovery of information and adjustment for dependent censoring using surrogate markers, in AIDS Epidemiology-Methodological Issues, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhauser, **1992**, pp. 297-331.

(5) Robins, J. M. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, in Proceedings of the Amer. Statist. Asso. Biopharmaceutical Section, Alexandria, VA: ASA, **1993**, pp. 24-33.

(6) Robins, J. M.;Finkelstein, D. Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted(IPCW) log-rank tests. Biometrices, **2000**, *56*, 779-788.

(7) Statten, G. A. and Datta S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. Amer. Statist. Ass., **2001**, *55*, 207-210.

(8) Lynden-Bell, D. A method of allowing for known observational selection in small samples applied to 3CR quasars. Mon. Not. R. Astr. Soc. **1971**, *155*, 95-118.

(9) Cox, D. R.; Oakes, D. Analysis of Survival Data. London: Chapman and Hall. **1984**.

(10) Woodroofe, M. Estimating a distribution function with truncated data. Ann. Statist., **1985**, *13*, 163-167.

(11) Wang, M.-C.; Jewell, N. P.; Tsai, W.-Y. Asymptotic properties of the product-limit estimate under random truncation. Ann. Statist., **1986**, *14* 1597-1605.

(12) Tsai, W.-Y.; Jewell, N. P.; Wang, M.-C. A note on the product-limit estimate under right censoring and left truncation. Biometrika, **1987**, *74*, 883-886.

(13) He, S.; Yang, G. L. Estimation of the truncation probability in the random truncation model. Ann. Statist., **1998**, *26*, 1011-1027.

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

(15) Vardi, Y. Empirical distribution in selection bias models. Ann. Statist., **1982**, *10*, 616-620.

(16) Turnbull, B. W. The empirical distribution with arbitrarily grouped, censored and truncated data. J. R. Statist. Soc. B., **1976**, *38*, 290-295.

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

(18) Gross, S. T.; Lai, T. L. Boostrap methods for truncated data and censored data. Statist. Sinica, **1996**, *6*, 509-530.

(19) Satten, G. A.; Datta, S.; Robins J. M. Estimating the marginal survival function in the presence of time dependent covariates. Statist. Prob. Lett., **2001**, *54*, 397-403.