東海大學統計研究所

碩士論文

指導教授:沈葆聖博士

雙截區間設限資料下無母數最大概似估計

The NPMLE for Doubly Truncated and Interval-censored Data

研究生:吳玠錡

中華民國一百零四年六月

論文致謝詞

此篇論文的完成，首先要感謝我的指導老師---沈葆聖教授，在論文的寫作過程中用心的指導與鼓勵，以及在論文完成的背後有著更多看不到的老師辛苦的付出。

此外也要感謝特地從台北前來東海大學口試本論文的戴政教授，在口試過程中給予了很多寶貴的意見，讓學生收獲良多。以及感謝系上的林正祥教授，在口試過程中不辭辛勞地審查與提出寶貴的意見。

感謝朋友對我的祝福，以及系上的同學互相的幫忙與照顧，最後感謝父母無盡的關懷與付出，和無限的支持與鼓勵。

The NPMLE for Doubly Truncated and Interval-censored Data

Director: Pao-sheng Shen
Student:Chieh-Chi Wu
Department of Statistics
Tunghai University
Taichung, Taiwan 40704

Abstract

In this note we consider nonparametric estimation with doubly truncated and interval-censored (DTIC) data. Using standard convex optimization techniques, we proposed verifiable necessary conditions for Turnbull (1976)'s estimator to be a nonparametric maximum likelihood estimator (NPMLE). We discuss the existence of the NPMLE based on conditions proposed by Hudgens (2005).

Keywords: nonparametric maximum likelihood estimator; self-consistency algorithm; interval-censoring; double truncation

Contents

Chapter 1

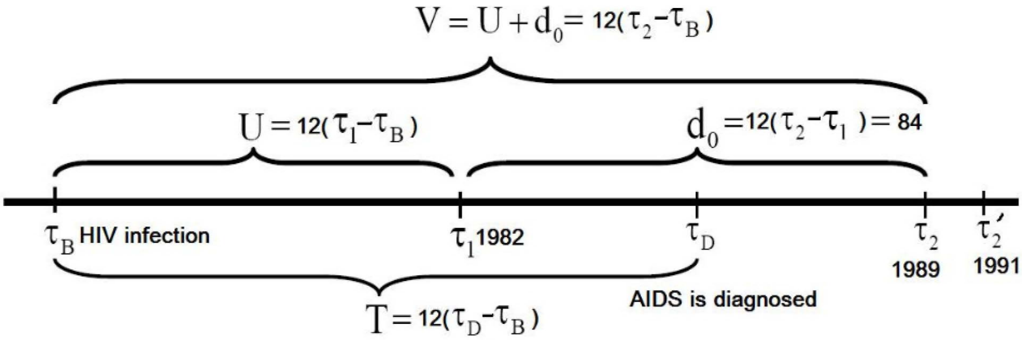
Introduction

Doubly truncated survival data arise when event times are observed only if they occur within subject specific intervals of times. This type of data play an important role in the statistical analysis of astronomical observations (see Efron and Petrosian (1999), Moreira and de Uña-Álvarez (2010), Shen (2010)) as well as in survival analysis (see Kalbfleisch and Lawless (1989), Bilker and Wang (1996)). Let T denote the failure time of interest and U and V denote the left-truncated and right-truncated variables. For doubly truncated data, one observes nothing if $T < U$ or $T > V$, and observes (T, U, V) if $U \leq T \leq V$. In many situations, the failure time is recorded in an interval $[L, R]$. Hence, T is subject to double truncation and interval censoring. For doubly truncated and interval-censored (DTIC) data, one observes nothing if $T < U$ or $T > V$, and observes (L, R, U, V) if $U \leq T \leq V$, where $[L, R] \subset [U, V]$. Consider the following application:

Example: 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. 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 ($\tau_2^{'}$), i.e. an HIV-infected population of interest. 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. Nevertheless, cases either diagnosed or reported after June 30, 1989 (τ_2), were not included (i.e. right truncated) to avoid bias resulting from reporting delay. Also, cases having the AIDS prior to July 1, 1982 (τ_1) 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_1 would have been missed (i.e. left-truncated). Let τ_B be the calendar time (in years) of the initiating events (HIV infection), and τ_D be the calendar time (in years) of AIDS onset. Let $T = 12(\tau_D - \tau_B)$ (in month) be the incubation time from HIV infection to AIDS. Let $U = 12(\tau_1 - \tau_B)$ (in month) and $V = 12(\tau_2 - \tau_B) = U + d_0$ (in month), where $d_0 = 12(\tau_2 - \tau_1) = 84$. Hence, T is observable only when $\tau_1 \leq \tau_D \leq \tau_2$ (i.e. $U \leq T \leq V$). Assume for each individual, data is available on a $p \times 1$ vector of covariates, $Z = [Z_1^*, \ldots, Z_p^*]^T$ (e.g. treatment, gender). It is important to investigate the association between Z and survival function of T. In this article, we will confine our attention to the situation where Z is discrete. Figure 1 highlights all the different times for doubly truncated data described in Example 1.

For arbitrarily censored and truncated data, Turnbull (1976) provided a self-consistency algorithm to obtain a nonparametric estimator for the distribution function of the failure time of interest. Frydman (1994) noted that the characterization given by Turnbull involves only censoring intervals and indicated that how Turnbull's characterization can be easily



$V = U + d_0 = 12(\tau_2 - \tau_B)$
$U = 12(\tau_1 - \tau_B)$
$d_0 = 12(\tau_2 - \tau_1) = 84$
τ_B HIV infection
τ_1 1982
τ_D
τ_2
1989
τ_2'
1991
AIDS is diagnosed
$T = 12(\tau_D - \tau_B)$

Figure 1. Schematic depiction of doubly truncated data

modified to obtain the appropriate characterization when data are subject to censoring and truncation. Alioum and Commenges (1996) identified a further refinement of the set where the nonparametric maximum likelihood estimate (NPMLE) could put mass. For interval-censored and left-truncated data, Hudgens (2005) proposed a graph theoretical approach to describe the support set of the NPMLE for the cumulative distribution function. Using EM algorithm of Turnbull (1976) and iterative convex minorant (ICM) algorithm (Groeneboom and Wellner (1992)), Shen (2012) studied the performance of the NPMLE of the distribution function of T. Simulation results indicate that the NPMLE performs adequately for finite sample. When there is no truncation, asymptotic properties of the NPMLE have been derived.(see Groeneboom and Wellner (1992), Gentleman and Geyer (1994), Yu et al. (1998a, 1998b), Shick and Yu (2000), van der Vaart and Wellner (2000), Song (2004)). However, much less is known about the large sample properties of the NPMLE if the data are subject to interval censoring and truncation.

In Section 2, we demonstrate that the Kuhn-Tucker conditions provide easily verifiable necessary conditions for Turnbull (1976)'s estimator to be a NPMLE. We discuss the existence of the NPMLE based on conditions proposed by Hudgens (2005). Several examples are given to illustrate the existence of the NPMLE. When there is no censoring, we show that Turnbull (1976)'s estimator is asymptotically equivalent to Efron and Petrosian (1999)'s estimator if the initial probabilities assigned at the leftmost and rightmost innermost sets tend to zero.

Chapter 2

The NPMLE

2.1 Interval-censored and Doubly-truncated Data

Let $(L_1, R_1, U_1, V_1), \ldots, (L_n, R_n, U_n, V_n)$ denote the observed DTIC data. Notice that $P([L_i, R_i] \subset [U_i, V_i]) = 1$. Without loss of generality, suppose the observed data are ordered according to L_i such that $L_1 \leq L_2 \leq \cdots \leq L_n$. Following Turnbull (1976) and Frydman (1994), we consider nonparametric estimation of $F(t)$ using the n independent pairs $\{A_1, B_1\}, \ldots, \{A_n, B_n\}$, where $A_i = [L_i, R_i]$ and $B_i = [U_i, V_i]$. Given B_i, the conditional likelihood of F is given by

$$L_c(F) = \prod_{i=1}^{n} \frac{P_F(A_i)}{P_F(B_i)},$$

where $P_F(R)$ is the probability assigned to the interval R by F. We define an NPMLE as $\tilde{F} = \text{argmax}_{F \in \mathcal{F}}\{L_c(F)\}$, where $\mathcal{F}$ denotes the class of distribution functions such that $P_F(\cup_{i=1}^{n} B_i) = 1$ and $L_c(F)$ is defined, i.e. $P_F(B_i) > 0$ for all $i = 1, \ldots, n$. Based on the approach of Hudgens (2005), we define $\mathcal{M} = \{M_1, M_2, \ldots, M_{3n}\}$, where $M_i = A_i$ for $i = 1, \ldots, n$, $M_i = [0, U_i]$ for $i = n+1, \ldots, 2n$ and $M_i = [V_i, \infty)$ for $i = 2n+1, \ldots, 3n$. Thus, we can obtain innermost intervals $H_j, j = 1, \ldots, J$, induced by $M_1, \ldots, M_{3n}$ to be all the disjoint intervals which are non-empty intersections of these M_i's such that $M_i \cap H_j = \emptyset$ or H_j for all i and j. We shall assume throughout that $H_1, \ldots, H_J$ are ordered and let the endpoints of the innermost intervals H_j be q_j and p_j, $j = 1, \ldots, J$, where

$$0 = q_1 \leq p_1 \leq q_2 \leq p_2 \leq \cdots \leq q_J \leq p_J = \infty.$$

Notice that the interval $[q_j, p_j]$ can be constructed (see Alioum and Commenges (1999)) by representing on the real line the elements of $\mathcal{L} = \{L_i : i = 1, \ldots, n\} \cup \{V_i : i = 1, \ldots, n\} \cup \{0\}$ and $\mathcal{R} = \{R_i : i = 1, \ldots, n\} \cup \{U_i : i = 1, \ldots, n\} \cup \{\infty\}$ by left hooks and right hooks, respectively. By going over the real line in the direct sense, the intervals $[q_j, p_j]$ are intervals opened by a left hook and closed by a right hook, and which contain no other hook. Similar to the Lemma 1 of Hudgens (2005), we can show that any distribution function which increases outside $\cup_{j=1}^{J} H_j$ cannot be an NPMLE of $L_c(F)$. Furthermore, for fixed value of $P_F(H_j)$, by inspection of the likelihood function $L_c(F)$, it follows that $L_c(F)$ is independent of the values of F within the region H_j. These conclusions allow us to consider maximizing the following simpler likelihood

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

where $\mathbf{s} = (s_1, \cdots, s_J)^T$, $s_j = P_F(H_j)$, $\alpha_{ij} = I_{[H_j \subset A_i]}$, and $\beta_{ij} = I_{[H_j \subset B_i]}$. 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 (2.3)-(2.5) as follows:

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

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

and

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

Thus, we can limit our search to the smaller space given by constraints (2.3)-(2.5). We shall use Ω to denote the parameter space that is given by constraints (2.3)-(2.5), i.e.

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

To find the maximum likelihood estimate of the vector $\mathbf{s}$, using an EM algorithm, we obtain Turnbull's (1976) self-consistency algorithm as follows:

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

where

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

and

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

By (2.6), for the b^{th} iteration, we have

$$\hat{s}_j^{(b)} - \hat{s}_j^{(b-1)} = \frac{\sum_{i=1}^{n} \alpha_{ij} \hat{s}_j^{(b-1)} / [\sum_{m=1}^{J} \alpha_{im} \hat{s}_m^{(b-1)}] - \sum_{i=1}^{n} \beta_{ij} \hat{s}_j^{(b-1)} / [\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}]}{\sum_{i=1}^{n} 1 / [\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}]}. \tag{2.7}$$

Next, under (2.2), the log likelihood of $L_c(\mathbf{s})$, denoted by $l(\mathbf{s})$, is the following:

$$l(\mathbf{s}) = \sum_{i=1}^{n} \left[\log\big(\sum_{j=1}^{J} \alpha_{ij} s_j\big) - \log\big(\sum_{j=1}^{J} \beta_{ij} s_j\big) \right].$$

To find the maximum estimate of the vector $\mathbf{s}$ we maximize $l(\mathbf{s})$ with respect to s subject to constraints (2.3) and (2.4). For a concave programming problem with linear constraints, the Kuhn-Tucker conditions (Gentleman and Geyer (1994)) are necessary and sufficient for optimality (Rockafellar (1970), Theorem 28.1, Corollary, 28.2.2). Let $\hat{s}_j$ $(j = 1, \ldots, J)$ denote the estimators obtained from (2.6).

Proposition 1.

The estimator $\hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_J]^T$ is the local maximizer if the following conditions are satisfied: (1) $\sum_{j=1}^{J} \hat{s}_j = 1$, (2) $\hat{s}_j \geq 0$ $(j = 1, \ldots, J)$ and (3) $d_j(\hat{\mathbf{s}}) = t_j(\hat{\mathbf{s}})$ for $\hat{s}_j > 0$; $t_j(\hat{\mathbf{s}}) \geq d_j(\hat{\mathbf{s}})$ for $\hat{s}_j = 0$, where

$$d_j(\hat{\mathbf{s}}) = \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{l=1}^{J} \alpha_{il} \hat{s}_l} \quad \text{and} \quad t_j(\hat{\mathbf{s}}) = \sum_{i=1}^{n} \frac{\beta_{ij}}{\sum_{l=1}^{J} \beta_{il} \hat{s}_l}.$$

Proof: A point $\hat{\mathbf{s}}$ is the local maximizer if and only if there exists Lagrange multipliers $\lambda_j(\hat{\mathbf{s}})$ $(j = 0, \ldots, J)$ such that the Kuhn-Tucker conditions (2.8)-(2.12) hold, with

$$\sum_{j=1}^{J} \hat{s}_j = 1, \tag{2.8}$$

$$\hat{s}_j \geq 0 \ (j = 1, \ldots, J), \tag{2.9}$$

$$\lambda_j(\hat{\mathbf{s}}) \hat{s}_j = 0 \ (j = 1, \ldots, J), \tag{2.10}$$

$$\lambda_j(\hat{\mathbf{s}}) \geq 0 \ (j = 1, \ldots, J), \tag{2.11}$$

$$\frac{\partial}{\partial s_j} \left\{ l_c(\mathbf{s}) + \sum_{j=1}^{J} s_j (\lambda_j - \lambda_0) \right\} \Bigg|_{s_j = \hat{s}_j} = d_j(\hat{\mathbf{s}}) - t_j(\hat{\mathbf{s}}) + \lambda_j(\hat{\mathbf{s}}) - \lambda_0(\hat{\mathbf{s}}) = 0 \ (j = 1, \ldots, J). \tag{2.12}$$

Multiplying (2.12) by $\hat{s}_j$ and summing yields $\lambda_0(\hat{\mathbf{s}}) = n - n = 0$. Hence, (2.12) is reduced to

$$d_j(\hat{\mathbf{s}}) - t_j(\hat{\mathbf{s}}) + \lambda_j(\hat{\mathbf{s}}) = 0. \tag{2.13}$$

If $\hat{s}_j > 0$ then (2.12) implies that $\lambda_j(\hat{\mathbf{s}}) = 0$, and (2.13) implies that $d_j(\hat{\mathbf{s}}) = t_j(\hat{\mathbf{s}})$. If $\hat{s}_j = 0$ then (2.11) implies that $\lambda_j(\hat{\mathbf{s}}) \geq 0$, and (2.12) implies that $t_j(\hat{\mathbf{s}}) \geq d_j(\hat{\mathbf{s}})$. Thus, for any $\hat{\mathbf{s}}$ that satisfies conditions (1)-(3), it is the NPMLE. The proof is complete.

Given j, let $\mathcal{A}_j = \{i : \alpha_{ij} = 1\}$ and $\mathcal{B}_j = \{i : \beta_{ij} = 1\}$. Since $\beta_{ij} = 1$ implies $\alpha_{ij} = 1$, $\mathcal{A}_j \subset \mathcal{B}_j$. Notice that the leftmost and rightmost innermost sets are equal to $[q_1, p_1] = [0, U_{(1)}]$ and $[q_J, p_J] = [V_{(n)}, \infty]$, respectively, where $U_{(1)}$ is the smallest variable of U_i's and $V_{(n)}$ is the largest variable of V_i's. Since $\mathcal{B}_1 = \mathcal{B}_J = \emptyset$, we have $d_j(\hat{\mathbf{s}}) = t_j(\hat{\mathbf{s}}) = 0$ for $j = 1, J$. Thus, condition (3) of Proposition 1 always holds and we can assign zero probabilities on these two sets, i.e. $\hat{s}_1 = \hat{s}_J = 0$. Next, we discuss two cases when the probabilities of the other innermost sets are equal to zero.

Case 1:

$$\textit{An innermost set } [q_j, p_j] \textit{ with } \beta_{ij} = 1, \ \alpha_{ij} = 0, \textit{ and } \beta_{i'j} = 0 \ \textit{for all } i' \neq i. \tag{2.14}$$

If an innermost satisfies condition (2.14), then it can only be covered by one truncation set and must has the form $[V_j, U_k]$. Under (2.14), $\mathcal{B}_j = \{i\}$ and $\mathcal{A}_j = \emptyset$, i.e. $\alpha_{ij} = 0$. By $d_j(\hat{\mathbf{s}}) = 0 < t_j(\hat{\mathbf{s}})$, we need to assign zero probability on $[q_j, p_j]$, i.e. $\hat{s}_j = 0$.

Case 2:

$$An\ innermost\ set\ [q_j, p_j]\ with\ \alpha_{ij} = 1,\ and\ \beta_{i^{'}j} = 0\ for\ all\ i^{'} \neq i. \tag{2.15}$$

If an innermost satisfies condition (2.15), then it can only be covered by two sets, one interval set and its corresponding truncation set. Under (2.15), we have $\mathcal{A}_j = \mathcal{B}_j = \{i\}$. By $d_j(\hat{\mathbf{s}}) = t_j(\hat{\mathbf{s}})$, we need to solve the equation $\sum_{l=1}^{J} \alpha_{il}\hat{s}_l = \sum_{l=1}^{J} \beta_{il}\hat{s}_l$. In this case, we need to assign zero probability on $[q_l, p_l]$ for any innermost set with $\alpha_{il} = 0$ and $\beta_{il} = 1$.

Thus, the NPMLE assign zero probabilities on innermost sets when either Case 1 or Case 2 occurs. The assignment of zero probability may cause the nonexistence of the NPMLE. Consider the following example.

Example 1

For example, for $n = 3$ with observations $U_1 < L_1 < U_2 < R_1 < V_1 < U_3 < L_2 < R_2 < V_2 < L_3 < R_3 < V_3$, we have innermost sets $[q_1, p_1] = [0, U_1]$, $[q_2, p_2] = [L_1, U_2]$, $[q_3, p_3] = [V_1, U_3]$, $[q_4, p_4] = [L_2, R_2]$, $[q_5, p_5] = [L_3, R_3]$, and $[q_6, p_6] = [V_3, \infty]$. In this example, $\alpha_{12} = \alpha_{24} = \alpha_{35} = 1$, $\beta_{12} = \beta_{23} = \beta_{24} = \beta_{34} = \beta_{35} = 1$, and the rest of indicators are equal to zero. Since $\beta_{23} = 1$, $\beta_{i^{'}3} = 0$ for $i^{'} = 1, 3$, and $\alpha_{23} = 0$, we have $d_3(\hat{\mathbf{s}}) = 0$ and $t_3(\hat{\mathbf{s}}) = 1/(\hat{s}_3 + \hat{s}_4)$. Thus, $\hat{s}_3 = 0$. Furthermore, since $\beta_{35} = \beta_{34} = 1$, $\beta_{i^{'}5} = 0$ for $i^{'} = 1, 2$, and $\alpha_{35} = 1$, we have $d_5(\hat{\mathbf{s}}) = 1/\hat{s}_5$ and $t_5(\hat{\mathbf{s}}) = 1/(\hat{s}_4 + \hat{s}_5)$. Thus, $\hat{s}_4 = 0$. However, since $d_4(\hat{\mathbf{s}}) = 1/\hat{s}_4$ and $t_4(\hat{\mathbf{s}}) = 1/(\hat{s}_3 + \hat{s}_4) + 1/(\hat{s}_4 + \hat{s}_5)$. Both $d_4(\hat{\mathbf{s}})$ and $t_4(\hat{\mathbf{s}})$ are infinity by setting $\hat{s}_3 = 0$ and $\hat{s}_4 = 0$. Thus, the NPMLE does not exist.

For left-truncated and interval-censored (LTIC) data, Hudgens (2005, Theorem 1) proposed a necessary and sufficient condition for the existence of an NPMLE, which can be rephrased as follows:

$$For\ each\ non{-}empty\ proper\ subset\ S\ of\ \mathcal{C} = \{1, \ldots, n\}, let\ S^c\ denote\ the\ complement\ of\ S.$$

$$There\ exists\ an\ i \in S^c\ such\ that\ if\ \alpha_{ij} = 1\ then\ \beta_{i^{'}j} = 1\ for\ some\ i^{'} \in S. \qquad (2.16).$$

Let S_{-i} be the subset of Ω with i deleted from Ω. Let $S_{-i}^c = \{i\}$ be the complement set of S_{-i}. Since the only element in S_{-i}^c is the element i, under (2.16), the following condition holds:

$$\text{if } \alpha_{ij} = 1,\ then\ there\ exists\ an\ i^{'} \neq i\ such\ that\ \beta_{i^{'}j} = 1. \tag{2.17}$$

Condition (2.17) requires that given an innermost set is covered by a censoring set $[L_i, R_i]$ then it must be covered by some truncation set $[U_{i^{'}}, V_{i^{'}}]$ with $i^{'} \neq i$. Under (2.17), Case 2 will not occur.

Hudgens (2005) pointed out that for DTIC data condition (2.16) is a sufficient condition for the existence of the NPMLE but not a necessary condition. Hudgens (2005) illustrated this point using the following example.

Example 2

For example, for $n = 3$ with $U_1 = U_2 < L_1 < U_3 < L_2 < R_1 < L_3 < R_3 < V_3 < R_2 < V_1 = V_2$. In this case, we have $[q_1, p_1] = [0, U_1]$, $[q_2, p_2] = [L_1, U_3]$, $[q_3, p_3] = [L_2, R_1]$, $[q_4, p_4] = [L_3, R_3]$, $[q_5, p_5] = [V_3, R_2]$ and $[q_6, p_6] = [V_1, \infty]$. Although condition (2.16) is not satisfied with the choice of $S = \{3\}$, the weaker condition (2.17) is satisfied, e.g. $\alpha_{12} = 1 \rightarrow \beta_{22} = 1$, $\alpha_{13} = 1 \rightarrow \beta_{23} = 1$, $\alpha_{11} = \alpha_{14} = \alpha_{15} = 0$, $\alpha_{21} = \alpha_{22} = 0$, $\alpha_{23} = 1 \rightarrow \beta_{13} = 1$, $\alpha_{24} = 1 \rightarrow \beta_{14} = 1$, $\alpha_{25} = 1 \rightarrow \beta_{15} = 1$, $\alpha_{31} = \alpha_{32} = \alpha_{33} = \alpha_{35} = 0$, and $\alpha_{34} = 1 \rightarrow \beta_{24} = 1$.

Moreover, $d_2(\hat{\mathbf{s}}) = t_2(\hat{\mathbf{s}}) \rightarrow 1/(\hat{s}_2 + \hat{s}_3) = 2$, i.e. $\hat{s}_2 + \hat{s}_3 = 0.5$; $d_3(\hat{\mathbf{s}}) = t_3(\hat{\mathbf{s}}) \rightarrow 1/(\hat{s}_2 + \hat{s}_3) + 1/(\hat{s}_3 + \hat{s}_4 + \hat{s}_5) = 2 + 1/(\hat{s}_3 + \hat{s}_4)$, i.e. $\hat{s}_5 = 0$; $d_4(\hat{\mathbf{s}}) = t_4(\hat{\mathbf{s}}) \rightarrow 1/(\hat{s}_3 + \hat{s}_4) + 1/\hat{s}_4 = 2 + 1/(\hat{s}_3 + \hat{s}_4)$, i.e. $\hat{s}_4 = 0.5$. Finally, since $\hat{s}_5 = 0$, we need $d_5(\hat{\mathbf{s}}) \leq t_5(\hat{\mathbf{s}}) \rightarrow 1/(\hat{s}_3 + \hat{s}_4 + \hat{s}_5) \leq 2$, i.e. $\hat{s}_3 + 0.5 \geq 0.5$, which always holds. Thus, for any $\delta \in [0, 0.5]$, the estimator $\hat{\mathbf{s}} = (0, 0.5 - \delta, \delta, 0.5, 0, 0)^T$ is the NPMLE.

The following example demonstrate that for DTIC data (2.17) is not a necessary condition for the existence of the NPMLE although it is a weaker condition than (2.16).

Example 3

For example, for $n = 2$ with $U_1 < L_1 < R_1 < V_1 < U_2 < L_2 < R_2 < V_2$. In this case, we have $[q_1, p_1] = [0, U_1]$, $[q_2, p_2] = [L_1, R_1]$, $[q_3, p_3] = [V_1, U_2]$, $[q_4, p_4] = [L_2, R_2]$ and $[q_5, p_5] = [V_2, \infty]$. In this example, $\alpha_{12} = \alpha_{24} = 1$, $\beta_{12} = \beta_{24} = 1$, and the rest of indicators are equal to zero. Since $\alpha_{12} = 1$, $\beta_{12} = 1$, and $\beta_{22} = 0$, (2.17) does not hold. For $\hat{s}_j > 0$, we have $d_2(\hat{\mathbf{s}}) = 1/\hat{s}_2 = t_2(\hat{\mathbf{s}})$ and $d_4(\hat{\mathbf{s}}) = 1/\hat{s}_4 = t_4(\hat{\mathbf{s}})$. Thus, for any $\delta \in [0, 1]$, the estimator $\hat{\mathbf{s}} = (0, \delta, 0, 1 - \delta, 0)^T$ is the NPMLE.

In conclusion, suppose there are K innermost sets with $\mathcal{A}_m = \emptyset$, then we assign zero probabilities on those K sets. For those innermost sets with nonempty $\mathcal{A}_j$, by $d_j(\hat{\mathbf{s}}) = t_j(\hat{\mathbf{s}})$, we need to solve $J - K$ equations as follows

$$\sum_{i \in \mathcal{A}_j} \frac{1}{\eta_i(\hat{\mathbf{s}})} = \sum_{i \in \mathcal{A}_j} \frac{1}{\zeta_i(\hat{\mathbf{s}})} + \sum_{i \in \mathcal{B}_j - \mathcal{A}_j} \frac{1}{\zeta_i(\hat{\mathbf{s}})}, \tag{2.18}$$

where $\eta_i(\hat{\mathbf{s}}) = \sum_{l=1}^{J} \alpha_{il} \hat{s}_l$ and $\zeta_i(\hat{\mathbf{s}}) = \sum_{l=1}^{J} \beta_{il} \hat{s}_l$. When $\mathcal{A}_j = \{i\}$, i.e. with only one element, Case 2 will not occur under (2.17) since $\mathcal{B}_j - \mathcal{A}_j \neq \emptyset$. If all the solutions are nonnegative, then the solution is the NPMLE. If some of $\hat{s}_j$'s are equal to zero, the solution is still the NPMLE if the left-hand side of (2.18) is smaller than right-side of (2.18) for those $\hat{s}_j$'s.

2.2. Doubly-truncated Data

In this section, we consider the special case when there is no censoring, i.e. doubly truncated data. Let $(T_1, U_1, V_1), \ldots, (T_n, U_n, V_n)$ denote the truncated sample, where $T_i = L_i = R_i$ is the truncated failure times. Let $\mathbf{f} = (f_1, \ldots, f_n)$ be a distribution putting probability f_i on T_i's. The conditional nonparametric likelihood can be written as

$$L_1(\mathbf{f}) = \prod_{j=1}^{n} \frac{f_j k_j}{\sum_{i=1}^{n} F_i k_i} = \prod_{j=1}^{n} \frac{f_j}{F_j},$$

where $F_i = \sum_{m=1}^{n} f_m J_{im}$, $J_{im} = I_{[U_i \leq T_m \leq V_i]} = 1$ if $U_i \leq T_m \leq V_i$ and equal to zero otherwise. Efron and Petrosian (1999) proposed a conditional NPMLE by maximizing the conditional log likelihood $l(\mathbf{f}) = \log(L_1(\mathbf{f}))$, which results in the following equation:

$$\hat{f}_j = \frac{1}{\sum_{i=1}^{n} J_{ij} / \sum_{m=1}^{n} \hat{f}_m J_{im}}. \quad (j = 1, \ldots, n) \tag{2.19}$$

An interesting question is " When there is no censoring, does Turnbull's estimator reduce to the Efron and Petrosian (1999)'s estimator ?" Intuitively, these two estimators should be different since Turnbull's estimator put probabilities at some innermost intervals which differs from T_i's. The following proposition shows that they are asymptoically equivalent to each other.

Proposition 2. If the initial probabilities assigned at $[0, U_{(1)}]$ and $[V_{(n)}, \infty]$ tend to zero, Turnbull (1976)'s estimator is asymptotically equivalent to Efron and Petrosian (1999)'s estimator.

Proof: For doubly truncated data, since $A_i = \{T_i\}$ (i.e. $T_i = E_i = R_i$), there are n innermost sets with a single point, i.e. $q_j = p_j$. Given j and $p_j = q_j$, we have $\alpha_{ij} = 1$ for some $i = i_j$ and equal to 0 for $i \neq i_j$. Thus, when $q_j = p_j$, since $\sum_{m=1}^{J} \alpha_{i_j m} \hat{s}_m^{(b-1)} = \hat{s}_j^{(b-1)}$, (2.7) is reduced to

$$\hat{s}_j^{(b)} - \hat{s}_j^{(b-1)} = \frac{1 - \sum_{i=1}^{n} \beta_{ij} \hat{s}_j^{(b-1)} / \left[\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}\right]}{\sum_{i=1}^{n} 1 / \left[\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}\right]},$$

i.e.

$$\hat{s}_j^{(b)} = \frac{1 + \sum_{i=1}^{n} (1 - \beta_{ij}) \hat{s}_j^{(b-1)} / \left[\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}\right]}{\sum_{i=1}^{n} 1 / \left[\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}\right]}. \tag{2.20}$$

Notice that (2.20) differs from (2.19).

Furthermore, for the other innermost sets with $p_j < q_j$, we have $\alpha_{ij} = 0$ for all i and (2.7) is reduced to

$$\hat{s}_j^{(b)} - \hat{s}_j^{(b-1)} = \frac{- \sum_{i=1}^{n} \beta_{ij} \hat{s}_j^{(b-1)} / \left[\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}\right]}{\sum_{i=1}^{n} 1 / \left[\sum_{m=1}^{J} \beta_{im} \hat{s}_m^{(b-1)}\right]}. \tag{2.21}$$

When $q_j < p_j$, since $\alpha_{ij} = 0$ for all i, we have $d_j(\hat{\mathbf{s}}) = 0 \leq t_j(\hat{\mathbf{s}})$, which implies that the mass $\hat{s}_j$ of the interval $[q_j, p_j]$ should be zero unless $\beta_{ij} = 0$ for all i. As $n \to \infty$, this can happen only at $[q_1, p_1] = [0, U_{(1)}]$ and $[q_J, p_J] = [V_{(n)}, \infty]$, and by (2.21), we have $\hat{s}_1^{(b)} - \hat{s}_1^{(b-1)} = 0$ and

$\hat{s}_J^{(b)} - \hat{s}_J^{(b-1)} = 0$. Thus, if the initial probabilities assigned at $[0, U_{(1)}]$ and $[V_{(n)}, \infty]$ tend to zero, Turnbull (1976)'s estimator asymptotically put probabilities only at T_i's. In this case, we have $J = n + 2$ and the estimator $\hat{s}_j^{(b)}$ asymptotically satisfies the following equation:

$$\hat{s}_j^{(b)} = \frac{1 + \sum_{i=1}^{n}(1 - J_{ij})\hat{s}_j^{(b-1)} / \left[\sum_{m=2}^{n+1} J_{im}\hat{s}_m^{(b-1)}\right]}{\sum_{i=1}^{n} 1 / \left[\sum_{m=2}^{n+1} J_{im}\hat{s}_m^{(b-1)}\right]},$$

which implies that

$$\hat{s}_j^{(b-1)} \sum_{i=1}^{n} J_{ij} \frac{1}{\sum_{m=2}^{n+1} J_{im}\hat{s}_m^{(b-1)}} = 1.$$

By (2.19), the proof is complete.

Chapter 3

Concluding Remarks

In this note, for ICDT data, we have proposed verifiable conditions for Turnbull (1976)'s estimator to be an NPMLE. We point out that when condition (2.17) does not hold, we need to assign zero probabilities on innermost sets, which may induce the nonexistence of the NPMLE. However, it is not a necessary condition although it is a weaker condition than (2.16). Further research is required to establish a necessary and sufficient condition for ICDT data.

References

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

Bilker, W. B. and Wnag, M.-C. (1996). A semiparametric extension oft the Mann-Whitney test for randomly truncated data. *Biometrics*, **52**, 10-20.

Efron, B. and Petrosian, V. (1999). Noparametric methods for doubly truncated data. *Journal of the American Statistical Association*, **94**, 824-834.

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.

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.

Moreira, C. and de Uña-Álvarez, J. (2010). A semiparametric estimator of survival for doubly truncated data. *Statistics in Medicine*

Rockafellar, R. Tyrrell (1970). Convex Analysis. Princeton, NJ: Princeton University Press. ISBN 978-0-691-01586-6

Shen, P-S. (2010). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics*, **62**, 5, 835-853.

Shen, P-S. (2012). Nonparametric analysis of interval censored and doubly truncated data. *Journal of Computational Statistics and Simulation*, **82**, 1845-1854.

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.

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

Wang, M.-C. (1987). Product-limit estimates: a generalized maximum likelihood study. *Communication in Statistics: Theory and Methods* **6**, 3117-3132.

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.

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

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