**Abstract**

Doubly truncated data appear in a number of applications, including astronomy and survival analysis. For double-truncated data, the lifetime $T$ is observable only when $U \leq T \leq V$, where $U$ and $V$ are the left-truncated and right-truncated time, respectively. In some situation, the lifetime $T$ also suffers interval censoring. Using EM algorithm of Turnbull (1976), we propose a nonparametric estimate of the distribution function of $T$. We show the consistency of the conditional nonparametric maximum likelihood estimate. Simulation results indicate that the proposed estimator performs adequately when truncation is not severe.

Key Words: double truncation; interval censoring; nonparametric maximum likelihood estimation.

# Contents

# Chapter 1

# Introduction

Doubly truncated failure-time arises if an individual is potentially observed only if its failure-time falls within a certain interval, unique to that individual. Doubly truncated data play an important role in the statistical analysis of astronomical observations (see Lynden-Bell (1971)) as well as in survival analysis. Consider the following applications:

**Example 1: Cohort-of-cases Data on Anti-diabetic Treatment**

For the period 1992-2003 the Odense Pharmaco-epidemiological Database (OPED) (see Støvring et al. (2003)) contains subject specific information on all prescriptions for subsidized medications redeemed at any pharmacy in the County of Fyn, as well as information on births, deaths and migration into and out of the County of Fyn. The tracking of individuals is based on the Civil Registration Number (CRN) which is assigned to all at birth or first immigration into Denmark. For each individual we identified all prescriptions of anti-diabetic agents in OPED. The anti-diabetic drugs are characterized by the first three characters of the so-called ATC-code being A10. Assume that the minimal and maximal observable age (in years) at diabetic onset before death is known and denoted by $\tau_0$ and $\tau_M$, respectively. Let $\tau_1 = 1992$ and $\tau_2 = 2003$. Define a population as the individuals who are born after the calendar time (in years) $\tau_1 - \tau_M$ (i.e. 1992-maximum age=year of birth for the oldest person), and before $\tau_2 - \tau_0$ (i.e. 2003-minimum age=year of birth for the youngest person) and will be diagnosed with the diabetic before death. For the individuals of the population defined above, let $\tau_B$ be the calendar time (in years) of the initiating events (birth), and $\tau_D$ be the calendar time (in years) at diabetic onset. Let $T = \tau_D - \tau_B$ be the age (in years) at diabetic onset. For the population defined above, let $U = \tau_1 - \tau_B$ and $V = \tau_2 - \tau_B = U + d_0$, where $d_0 = \tau_2 - \tau_1$. Note that $U$ and $V$ denote the age (in years) at $\tau_1$ and $\tau_2$, respectively. Figure 1 highlights all the different times for doubly truncated data.

In terms of the population defined above, a selection bias (double-truncated) results from the exclusion of those individuals who develop diabetic before $\tau_1 = 1992$ (the age at onset

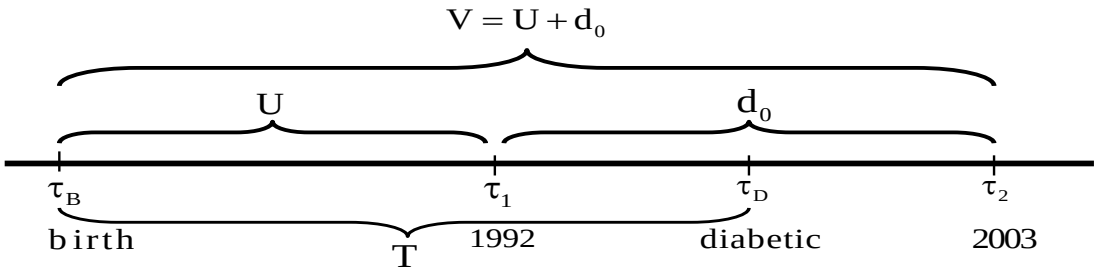


Figure 1. Schematic depiction of doubly truncated data

is too early) or after $\tau_2 = 2003$ (the age at onset is too late). Hence, $T$ is observable only when $U \leq T \leq V$.

**Example 2: 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 ($\tau_2$). 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 ($\tau_2$), were not included (i.e. right truncated) to avoid bias resulting from reporting delay. Also, cases having the AIDS prior to July 1, 1982 ($\tau_1$) 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 $\tau_1$ would have been missed (i.e. left-truncated). Let $\tau_B$ be the calendar time (in years) of the initiating events (HIV infection), and $\tau_D$ be the calendar time (in years) at which AIDS is diagnosed. 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 $\tau_1 \leq \tau_D + T/12 \leq \tau_2$ (i.e. $U \leq T \leq V$).

In Example 1, when the age of onset, $\tau_D$, is recorded exactly, we observe a doubly truncated sample $(T_1, U_1, V_1), \ldots, (T_n, U_n, V_n)$. Similarly, in Example 2, when the infection time $\tau_B$ and the time of onset of AIDS are both recorded exactly, we also observe a doubly truncated sample. Without loss of generality, suppose the observed data are ordered according to $T_i$ such that $T_1 < T_2 < \cdots < T_n$. Let $F(t)$ denote the distribution function of $T$, and $K(x, y)$ denote the bivariate distribution function of $(U, V)$. For any distribution function $W$ denote the left and right endpoints of its support by $a_W = inf\{t : W(t) > 0\}$ and $b_W = inf\{t : W(t) = 1\}$, respectively. Let $G(u) = K(u, \infty)$ and $Q(v) = K(\infty, v)$ be the

marginal distribution function of $U$ and $V$, respectively. Throughout this article we assume that

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

Under this assumption, $F$ and $K$ are both identifiable (see Woodroofe (1985)). For the special case, $V = U + d_0$, the assumption (1.1) is reduced to

$$a_G \leq a_F \leq a_G + d_0 \text{ and } b_G \leq b_F \leq b_G + d_0.$$

For doubly truncated data, the NPMLE of $F$ was first studied by Efron and Petrosian (1999)). The asymptotically properties were formally established by Shen (2008). The NPMLE is a discrete distribution putting all of its probability on the observed responses $(U_1, V_1, T_1), \ldots, (U_n, V_n, T_n)$. Let $\mathbf{f} = (f_1, \ldots, f_n)$ be a distribution putting probability $f_j$ on $T_j$ $(j = 1, \ldots, n)$. Similarly, let $\mathbf{k} = (k_1, \ldots, k_n)$ be a distribution putting joint probability $k_j$ on $(U_j, V_j)$ $(j = 1, \ldots, n)$. Under the assumption of independence of $T$ and $(U, V)$, the full nonparametric likelihood based on observed data can be written as

$$L(F) = \prod_{j=1}^{n} \frac{f_j k_j}{\sum_{i=1}^{n} F_i k_i} = \prod_{j=1}^{n} \frac{f_j}{F_j} \times \prod_{j=1}^{n} \frac{F_j k_j}{\sum_{i=1}^{n} F_i k_i} = L_1(\mathbf{f}) \times L_2(\mathbf{f}, \mathbf{k}),$$

where $F_i = \sum_{m=1}^{n} f_m J_{im}$, where $J_{im} = I_{[U_i \leq T_m \leq V_i]} = 1$ if $U_i \leq T_m \leq V_i$ and equal to zero otherwise. According to $L_1(\mathbf{f})$, the NPMLE of $\mathbf{f}$ can be obtained (see Efron and Petrosian (1999), Shen (2008)) by solving the following equation:

$$\frac{1}{\hat{f}_j} = \sum_{i=1}^{n} J_{ij} \frac{1}{\hat{F}_i}, \quad (j = 1, \ldots, n) \tag{1.2}$$

where $\hat{F}_i = \sum_{m=1}^{n} \hat{f}_m J_{im}$.

However, there are many applications, in which the age of onset $\tau_D$ suffers interval censoring, e.g. the onset of diabetes (or AIDS) is recorded only between an interval. Hence, the variable of interest $T$ is only recorded between an interval, say $[E, R]$. When there is no truncation, the first work on NPMLE of the $F$ in the presence of interval-censored data is attributed to Peto (1973). Turnbull (1976) characterized the NPMLE in the presence of interval censoring and truncation. Frydman (1994) later corrected Turnbull's characterization. Alioum and Commenges (1996) identified a further refinement of the set where the NPMLE can put mass. For left-truncated and interval-censored data, using a graph theoretical approach, Hudgens (2005) proposed a necessary and sufficient condition for the existence of the NPMLE. In Section 2, we propose an NPMLE of $F$ (denoted by $\hat{F}$) under interval-censoring and double truncation. Simulation results indicate that the estimator $\hat{F}$ performs adequately when truncation is not severe. Furthermore, under certain assumptions on support of $T$, $(U, V)$ and $(E, R)$, we show the consistency of the conditional NPMLE.

# Chapter 2

# The NPMLE

## 2.1 EM algorithm

When $T$ is subject to interval censoring, we only observe $[E, R] \subset [U, V]$ when $U \leq T \leq V$. Let $(E_1, R_1, U_1, V_1), \ldots, (E_n, R_n, U_n, V_n)$ denote the doubly truncated and interval-censored data. Without loss of generality, suppose the observed data are ordered according to $E_i$ such that $E_1 < E_2 < \cdots < E_n$. Following Turnbull (1976), Frydman (1994) and Alioum and Commenges (1996), we consider nonparametric estimation of $F$ using the $n$ independent pairs $\{A_1, B_1\}, \ldots, \{A_n, B_n\}$, where $A_i = [E_i, R_i]$ and $B_i = [U_i, V_i]$. Given $B_i$, the conditional likelihood of $F$ is given by

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

where $P_F(R)$ denotes the probability that is assigned to the interval by $F$. We define an NPMLE as $\hat{F} = \text{argmax}_{F \in \mathcal{F}}\{L_c(F)\}$, 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 $i = 1, \ldots, n$. Using the approach of Hudgens (2005), we define $\mathcal{K} = \{K_1, K_2, \ldots, K_{3n}\}$, where $K_1 = A_i$ for $i = 1, \ldots, n$, $K_i = (-\infty, U_i]$ for $i = n+1, \ldots, 2n$ and $K_i = [V_i, \infty)$ for $i = 2n+1, \ldots, 3n$. An intersection graph for $\mathcal{K}$ is constructed as follows. For each element of $\mathcal{K}$, we define a corresponding vertex. Let $i$ be the label of the vertex corresponding to $K_i$. Denote the set of vertex by $S_v$. Two vertices in $S_v$ are considered connected by an edge if and only if the two corresponding regions in $\mathcal{K}$ intersect. A clique is defined as a subset $M$ of $S_v$ such that every member of $M$ is connected by an edge to every other member of $M$. A maximal clique has the additional property that it is not a proper subset of any other clique. Let $\mathcal{M} = \{M_1, \ldots, M_J\}$ be the subset of maximal cliques of $S_v$ such that for each $M_j \in \mathcal{M}$, there is some $i \in \{1, \ldots, n\}$ such that $i \in M_j$. Let $\mathcal{H} = \{H_1, \ldots, H_J\}$ be the corresponding set of real representations of elements of $\mathcal{M}$ where $H_j = \cap_{i \in M_j} K_i$ for $j = 1, \ldots, J$. By Lemma 1 of Hudgens (2005), any distribution function which increases outside $\cup_{j=1}^{J} H_j$ cannot be

an NPMLE. By Lemma 2 of Hudgens (2005), for fixed value of $P_F(H_j)$, the likelihood is independent of the values of $F$ within the region $H_j$. These lemmas allow us to consider maximizing a simpler likelihood than equation (2.1). For each $H_j \in \mathcal{H}$, let $s_j = P_F(H_j)$ and let $\mathbf{s}$ be an m-dimension column vector with elements $s_j$. We shall assume throughout that $H_1, \ldots, H_J$ are ordered such that $H_j = [q_j, p_j]$ is to the left of $H_{j+1} = [q_{j+1}, p_{j+1}]$ for $j = 1, \ldots, J-1$, i.e. $[q_1, p_1]$, $[q_2, p_2], \ldots, [q_J, p_J]$, where $q_1 \leq p_1 < q_2 \leq p_2 < \cdots < q_J \leq p_J$. It follows that from lemmas 1 and 2 of Hudgens (2005) that maximizing likelihood (2.1) is equivalent to maximizing

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

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

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

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

and

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

Note that constraint (2.5) ensures that the likelihood is defined over the entire parameter space and this could instead be accomplished by the constraint

$$\sum_{j=1}^{J} \beta_{ij} s_j > 0, \ (i = 1, \ldots, n). \tag{2.6}$$

However, any $\mathbf{s}$ which satisfies constraints (2.3), (2.4) and (2.6), but not (2.5), will not maximize $L_c(\mathbf{s})$. Thus, we can limit our search to the smaller space given by constraints (2.3)-(2.5). We shall use $\Omega$ 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 $\mathbf{s}$, we use an EM algorithm as follows.

**E-Step**: Let the expected value of $\alpha_{ij}$ be denoted by $\mu_{ij}(\mathbf{s})$. Then under $\mathbf{s}$,

$$\mu_{ij}(\mathbf{s}) = \frac{\alpha_{ij} s_j}{\sum_{k=1}^{J} \alpha_{ik} s_k}. \tag{2.7}$$

Because of double truncation, the $i^{th}$ observation can be thought of as representing a group of unknown size where all observations in that group are unobserved because their failure time lies outside of $B_i$. We can refer to these observations as doubly truncated individuals. Let $\mathcal{G}_{ij}$ be the number in that group corresponding to the $i^{th}$ observation having failure in $H_j = [q_j, p_j]$. Note that the expected number of truncated individuals for the $i^{th}$ observation is $P(B_i^c)/P(B_i)$ and the probability of one of these truncated individuals falling in $H_j$ is $(1-\beta_{ij})s_j/P(B_i^c) = (1-\beta_{ij})s_j/[1-\sum_{k=1}^{n}\beta_{ik}s_k]$. Let the expected value of $\mathcal{G}_{ij}$ be denoted by $\eta_{ij}(\mathbf{s})$. Then

$$\eta_{ij}(\mathbf{s}) = [P(B_i^c)/P(B_i)] \times (1-\beta_{ij})s_j/P(B_i^c) = \frac{(1-\beta_{ij})s_j}{\sum_{k=1}^{J}\beta_{ik}s_k}. \tag{2.8}$$

**M-Step**: In the maximization step, we treat expected values as observed. The overall proportion of failures in the interval $H_j$ is

$$\pi_j(\mathbf{s}) = \frac{\sum_{i=1}^{n}[\mu_{ij}+\eta_{ij}]}{\sum_{i=1}^{n}\sum_{k=1}^{J}[\mu_{ik}+\eta_{ik}]}. \tag{2.9}$$

The EM algorithm iterates between equations (2.7), (2.8) and (2.9) after selecting initial estimates $s_j^{(0)} > 0$ such that $\sum_{j=1}^{J} s_j^{(0)} = 1$, i.e., computes $\mu_{ij}(\mathbf{s}^{(0)})$ and $\eta_{ij}(\mathbf{s}^{(0)})$, updates $\mathbf{s}$ by $\mu_{ij}(\mathbf{s}^{(1)})$ and $\eta_{ij}(\mathbf{s}^{(1)})$, and repeats until convergence. The resulting self-consistent estimate of $\mathbf{s}$, which is a solution of simultaneous equation $s_j = \pi_j(\mathbf{s})$ $(j = 1, \ldots, J)$, is exactly the Turnbull's (1976) self-consistency algorithm as follows:

$$s_j^{(b)} = \left\{1 + \frac{d_j(s^{(b-1)})}{M(s^{(b-1)})}\right\} s_j^{(b-1)} \ (1 \le j \le J), \tag{2.10}$$

where

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

and

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

Let $\hat{s}_j$ $(j = 1, \ldots, J)$ denote the estimators obtained from (2.10). As pointed out by Hudgens (2005), in general, a maximizer of $L_c(\mathbf{s})$ subject to $s \in \Omega$ need not exist since $\Omega$ is not closed. For left-truncated and interval-censored data, Hudgens (2005) (see Theorem 1, page 578) proposed a sufficient and necessary condition for the existence of the NMPLE as follows:

" There is a maximizer of $L_c(\mathbf{s})$ subject to $\mathbf{s} \in \Omega$ if and only if for each non-empty proper subset $S$ of $\{1, \ldots, n\}$ there is an $i \notin S$ such that $\mathcal{A}_i \subset \mathcal{D}_S$, $\mathcal{A}_i = \cup_{j \in A_i^*} H_j$, $\mathcal{D}_S = \cup_{k \in S}\mathcal{B}_k$, $\mathcal{B}_k = \cup_{j \in B_k^*} H_j$, where $A_i^* = \{j : \alpha_{ij} = 1\}$ and $B_k^* = \{j : \beta_{kj} = 1\}$".

For doubly truncated interval censored data (see Example 5 of Hudgens (2005)), the existence condition remains sufficient and is no longer necessary. In practice, when $n$ is large, it is difficult computationally to check the sufficient condition.

Alternatively, we can first determine that we have attained a local maximum by checking the necessary Kuhn-Tucker conditions (see Gentleman and Geyer (1994)) as follows: If $L_c(\mathbf{s})$ attains a local maximum then there exists Lagrange multiplies $\lambda_j$ $(j = 0, \ldots, J)$ such that $\sum_{j=1} s_j = 1$, $s_j \geq 0$ and (2.11)-(2.13) hold, with

$$\lambda_j s_j = 0 \ (j = 1, \ldots, J), \tag{2.11}$$

$$s_j \geq 0 \ (j = 1, \ldots, J), \tag{2.12}$$

$$\frac{\partial}{\partial s_j}\left\{ logL_c(\mathbf{s}) + \sum_{j=1}^{J} s_j(\lambda_j - \lambda_0) \right\} = d_j - t_j + \lambda_j - \lambda_0 = 0 \ (j = 1, \ldots, J), \tag{2.13}$$

where

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{k=1}^{J} \alpha_{ik} s_k} \quad \text{and} \quad t_j = \sum_{i=1}^{n} \frac{\beta_{ij}}{\sum_{k=1}^{J} \beta_{ik} s_k}.$$

Multiplying (2.13) by $s_j$ and summing yields $\lambda_0 = n - n = 0$. If $s_j > 0$ then (2.11) implies that $\lambda_j = 0$, and (2.13) implies that $d_j = t_j$. Conversely, if $s_j = 0$ then (2.13) implies that $d_j = t_j - \lambda_j$. Hence, when $s_j = 0$, $d_j$ can be larger or smaller than $t_j$. In practice, using the values of $s_j$, $d_j$ and $t_j$, we may first check if a local maximum estimator $\hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_J]^T$ of $\mathbf{s}$ is obtained. However, if the estimate is not unique, it may be that some of the masses are not identifiable. For example, it is possible that some $j \in \{1, \ldots, J\}$ and $j' \in \{1, \ldots, J\}$, the term $s_j$ and $s_{j'}$ appear in likelihood only in sum. This occurs if $\alpha_{ij} = \alpha_{ij'}$ and $\beta_{ij} = \beta_{ij'}$ for all $i$, in which case, only $s_j + s_{j'}$ is identifiable. Hence, inspection of $\sum_{i=1}^{n} \alpha_{ij}$ and $\sum_{i=1}^{n} \beta_{ij}$ for different $j$ is helpful in determining possible unidentifiable parameters. Another method for finding unidentifiable parameters is to look for estimates $\hat{\mathbf{s}}$ that vary for different initial starting values of the EM.

Based on the the estimator $\hat{s} = [\hat{s}_1, \ldots, s_J]^T$, an estimator $\hat{F}(t)$ of $F(t)$ can be uniquely defined for $t \in [p_j, q_{j+1})$ by $\hat{F}(p_j) = \hat{F}(q_{j+1}-) = \hat{s}_1 + \cdots + \hat{s}_j$, but is not uniquely defined for $t$ being in an open innermost interval. To avoid ambiguity we define $\hat{F}(t) = \hat{s}_1 + \cdots + \hat{s}_{j-1} + \hat{s}_j \frac{t-q_j}{p_j - q_j}$ if $t \in (q_j, p_j]$, $q_j > 0$ and $p_j \neq \infty$; $\hat{F}(t) = \hat{s}_1$ if $0 = q_1 \leq t \in [0, p_1]$; $\hat{F}(t) = 1 - \hat{s}_J$ if $t \geq q_J$ and $q_J < p_J = \infty$.

The following proposition shows that when all the widths of $R_i - E_i$ $(i = 1, \ldots, n)$ are small such that $\alpha_{ij} = 1$ for $i = j$ and $\alpha_{ij} = 0$ for $i \neq j$, equation (2.10) reduces to equation (1.2).

**Proposition 1.**

Suppose $\alpha_{ij} = 1$ for $i = j$ and $\alpha_{ij} = 0$ for $i \neq j$. Then equation (2.10) reduces to equation (1.2).

**Proof:**

Since $\alpha_{ij} = 1$ for $i = j$ and $\alpha_{ij} = 0$ for $i \neq j$, we have $\mu_{ij}(s) = 1$ for $i = j$ and $\mu_{ij}(s) = 0$ for $i \neq j$ and

$$d_j(s^{(b-1)}) = \frac{1}{s_j^{(b-1)}} - \sum_{i=1}^{n} \frac{\beta_{ij}}{\sum_{k=1}^{n} \beta_{ik} s_k^{(b-1)}}.$$

Hence, (2.10) reduces to

$$s_j^{(b)} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{n} \beta_{ik} s_k^{(b-1)}}} + s_j^{(b-1)} \left( \frac{\sum_{i=1}^{n} (1 - \beta_{ij}) / (\sum_{k=1}^{n} \beta_{ik} s_k^{(b-1)})}{\sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{n} \beta_{ik} s_k^{(b-1)}}} \right),$$

and

$$s_j^{(b)} - s_j^{(b-1)} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{n} \beta_{ik} s_k^{(b-1)}}} - s_j^{(b-1)} \left( \frac{\sum_{i=1}^{n} \beta_{ij} / (\sum_{k=1}^{n} \beta_{ik} s_k^{(b-1)})}{\sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{n} \beta_{ik} s_k^{(b-1)}}} \right). \tag{2.14}$$

By (2.14), it follows that $\hat{s}_j$ $(j = 1, \ldots, J)$ satisfies the following equation:

$$\frac{1}{\hat{s}_j} = \sum_{i=1}^{n} \frac{\beta_{ij}}{\sum_{k=1}^{n} \beta_{ik} \hat{s}_k}. \tag{2.15}$$

Denote $\beta_{ij}$ as $J_{ij}$. Then equation (2.15) reduces to equation (1.2). The proof is completed.

Notice that when $E_i = R_i$ $(i = 1, \ldots, n)$ (i.e. doubly truncated data), by Proposition 1, we have $\hat{s}_j = \hat{f}_j$ $(j = 1, \ldots, n)$.

## 2.2 Consistency of the NPMLE

When there is no truncation, asymptotic properties of the NPMLE have been derived for interval-censored data. Groeneboom and Wellner (1992) proposed an iterative convex minorant algorithm to calculate the NPMLE and proved the uniform consistency of the NPMLE when $F$ is continuous and the joint distribution function of $(E, R)$ is absolutely continuous. If $(E, R)$ is assumed discrete, the NPMLE has the usual $\sqrt{n}$ convergence rate and a normal limiting distribution (Yu et al. (1998a, b)). However, if $(E, R)$ is continuous, the NPMLE converges slower than $\sqrt{n}$ to a non-Gaussian limiting distribution (see Groeneboom and Wellner (1992), Shick and Yu (2000), van der Vaart and Wellner (2000), Song (2004)). Although asymptotic properties of the NPMLE have been derived for the interval-censored data without truncation much less is known about the large sample properties of the NPMLE if both interval censoring and truncation are present. Pan and Chappell (1999) showed that the NPMLE is inconsistent when data is subject to case 1 interval censoring and left truncation. Under the assumption of monotonic hazard function, Pan et al. (1998) showed the

consistency of the NPMLE when data is subject to left truncation and interval censoring. The following Theorem establishes the consistency of the conditional NPMLE.

**Theorem 1.**

Assume that the joint distribution function of $(E, R)$ (denoted by $F_I$) has a density $f_I$ satisfying $f_I(e, r) > 0$ for any $(e, r) \subset (a_F, b_F)$. Under assumption (1.1), with probability one for each vaguely convergent subsequence of the NPMLEs $\{\hat{F}\}$, its limit $F_*$ satisfies

$$\frac{F(r) - F(e)}{F(v) - F(u)} = \frac{F_*(r) - F_*(e)}{F_*(v) - F_*(u)} \quad L - a.s.$$

where $L$ is the joint distribution function of $(E, R, U, V)$.

**Proof:**

The likelihood for the $i^{th}$ the observation $y_i = (u_i, v_i, e_i, r_i)$ is

$$p(y_i; F) = \frac{F(r_i) - F(e_i)}{F(v_i) - F(u_i)}.$$

Since the NPMLE $\hat{F}$ maximizes the likelihood function $\prod_{i=1}^{n} p(y_i; F)$, we have

$$\sum_{i=1}^{n} \log p(y_i; \hat{F}) \geq \sum_{i=1}^{n} \log p(y_i; F),$$

and then

$$\sum_{i=1}^{n} \log \frac{p(y_i; \hat{F})}{p(y_i; F)} \geq 0.$$

By the concavity of the function $\log(x)$ and by Jensen's inequality, for any $0 < \alpha < 1$ and $x > 0$, we have

$$\log(1 - \alpha + \alpha x) \geq (1 - \alpha) \log 1 + \alpha \log x = \alpha \log x,$$

hence

$$\begin{aligned} \int \log\left(1 - \alpha + \alpha \frac{p(y; \hat{F})}{p(y; F)}\right) dP_n(y) &= n^{-1} \sum_{i=1}^{n} \log\left(1 - \alpha + \alpha \frac{p(y_i; \hat{F})}{p(y_i; F)}\right) \\ &\geq n^{-1} \sum_{i=1}^{n} \alpha \log \frac{p(y_i; \hat{F})}{p(y_i; F)} \geq 0, \end{aligned} \tag{2.16}$$

where $P_n$ is the empirical measure of $(U_i, V_i, E_i, R_i)$ $i = 1, \ldots, n$.

Let $P$ be the probability measure of $(U_i, V_i, E_i, R_i)$. The left hand side of (2.16) is

$$\int \log\left(1 - \alpha + \alpha \frac{p(y; \hat{F})}{p(y; F)}\right) d(P_n - P)(y) + \int \log\left(1 - \alpha + \alpha \frac{p(y; \hat{F})}{p(y; F)}\right) dP(y). \tag{2.17}$$

Now we take our sample space $\Omega_s$ to be the space of all infinite sequences

$(U_1, V_1, E_1, R_1)$,$(U_2, V_2, E_2, R_2)$,$\cdots$,endowed with the Borel $\sigma$-algebra generated by the product topology on $\prod_1^\infty R_+^4$ and the product measure $\mathbf{P}$. By corollary 8.1 of Huang and Wellner (1995) and the generalized Glivenko-Cantelli theorem, there is a $B \in \Omega_s$ with $\mathbf{P}(B) = 1$ such that for each $\omega \in B$ the first term in (2.17) converges to 0. Fix an $\omega \in B$. By Helly's selection theorem, for any subsequence of $\hat{F}(\omega)$, there exists a further subsequence $\hat{F}_{n_k}(w)$ converging vaguely to some nondecreasing function $F_*$ taking values in $[0, 1]$. Under assumption (1.1), $p(y; \hat{F}_{n_k})/p(y; F)$ is bounded. By the bounded convergence theorem,

$$\lim_{n_k \to \infty} \int \log\left(1 - \alpha + \alpha \frac{p(y; \hat{F}_{n_k})}{p(y; F)}\right) dP(y) = \int \log\left(1 - \alpha + \alpha \frac{p(y; F_*)}{p(y; F)}\right) dP(y).$$

By (2.16), the above expression must be nonnegative. However, by Jensen's inequality it must be non-positive. Therefore, it must be zero, which leads to

$$p(y; F_*) = p(y; F) \quad P - a.s.$$

and it implies that

$$\frac{F(r) - F(e)}{F(v) - F(u)} = \frac{F_*(r) - F_*(e)}{F_*(v) - F_*(u)} \quad L - a.s.$$

So it is shown that with probability one, for each vaguely convergent subsequence of $\hat{F}(\omega)$, its limit $F_*$ satisfies

$$\frac{F(r) - F(e)}{F(v) - F(u)} = \frac{F_*(r) - F_*(e)}{F_*(v) - F_*(u)}.$$

The proof is completed.

Although we only establish consistency of the conditional NPMEL. Simulation results in the following section indicate that the NPMLE performs adequately when truncation is not severe.

# Chapter 3

# Simulation Results

A simulation study is conducted to investigate the performance of the proposed estimator $\hat{F}(t)$. The $T$'s are i.i.d. exponential distributed with mean equal to 1. The $U$'s are i.i.d. exponential distributed with scale parameters $\theta = 2, 4$ and 8, i.e. $G(x;\theta) = 1 - exp(-\theta x)$ for $x > 0$. The $V$ is set as $V = U + 2.0$. The $T$ and $(U, V)$ are independent to each other. The goal is to estimate $F(t_p) = p$, with $p = 0.2$, 0.5 and 0.8. To make $T$ interval-censored, we generate a uniform random variable $X$. If $X \leq 0.5$ then $E = T - c$ and $R = T + c + 0.1$. If $X > 0.5$ then $E = T - (c + 0.1)$ and $R = T + c$. The values of $c$ are set at $c = 0.15, 0.25$. The sample sizes are chosen as 200 and 400. The replication is 1000 times. Tables 1 through 3 show the empirical biases, standard deviations (std.) and mean squared errors (mse) of $\hat{F}$. Tables 1 through 3 also list the proportion of truncation $1 - P(U < T < V)$ (denoted by $q_T$). Furthermore, we also consider the estimation of $F_C(t_p) = [F(t_p) - F(t_{0.1})]/[F(t_{0.9}) - F(t_{0.1})]$. Tables 1 through 3 show the empirical biases, standard deviations (std.) and mean squared errors (mse) of $\hat{F}_C(t_p) = [\hat{F}(t_p) - \hat{F}(t_{0.1})]/[\hat{F}(t_{0.9}) - \hat{F}(t_{0.1})]$ with $p = 0.2$, 0.5 and 0.8. Based on the results of Tables 1 through 3, we conclude that:

(i) Given $q_T$, the mse of both estimators $\hat{F}$ and $\hat{F}_C$ increase as the length of censoring (i.e. $c$) increases.

(ii) Given $c$, the mse of both estimators $\hat{F}$ and $\hat{F}_C$ increase as proportion of truncation $q_T$ increases.

(iii) Given $c$ and $q_T$, the mse of both estimators $\hat{F}$ and $\hat{F}_C$ decrease as sample size $n$ increases.
(iv) When truncation is severe (i.e. $q_T = 0.43$), the biases of the estimator $\hat{F}_C$ are much smaller than that of $\hat{F}$.

Table 1. Simulation results for bias, standard deviation and root mean squared error of $\hat{F}(t_{0.2})$ and $\hat{F}_C(t_{0.2})$

| | | | | $\hat{F}(t_{0.2})$ | | | $\hat{F}_C(t_{0.2})$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $c$ | $n$ | $q_T$ | bias | std | mse | bias | std | mse |
| 2.0 | 0.15 | 200 | 0.43 | -0.123 | 0.023 | 0.503 | -0.065 | 0.021 | 0.399 |
| 2.0 | 0.15 | 400 | 0.43 | -0.124 | 0.017 | 0.482 | -0.066 | 0.014 | 0.376 |
| 2.0 | 0.25 | 200 | 0.43 | -0.137 | 0.026 | 0.533 | -0.077 | 0.023 | 0.429 |
| 2.0 | 0.25 | 400 | 0.43 | -0.129 | 0.015 | 0.484 | -0.072 | 0.013 | 0.381 |
| 4.0 | 0.15 | 200 | 0.31 | -0.101 | 0.032 | 0.496 | -0.054 | 0.023 | 0.385 |
| 4.0 | 0.15 | 400 | 0.31 | -0.096 | 0.017 | 0.439 | -0.048 | 0.015 | 0.340 |
| 4.0 | 0.25 | 200 | 0.31 | -0.129 | 0.037 | 0.550 | -0.076 | 0.025 | 0.434 |
| 4.0 | 0.25 | 400 | 0.31 | -0.107 | 0.015 | 0.448 | -0.059 | 0.012 | 0.350 |
| 8.0 | 0.15 | 200 | 0.23 | -0.098 | 0.035 | 0.499 | -0.052 | 0.023 | 0.380 |
| 8.0 | 0.15 | 400 | 0.23 | -0.074 | 0.022 | 0.421 | -0.040 | 0.015 | 0.323 |
| 8.0 | 0.25 | 200 | 0.23 | -0.137 | 0.040 | 0.570 | -0.086 | 0.026 | 0.455 |
| 8.0 | 0.25 | 400 | 0.23 | -0.098 | 0.026 | 0.474 | -0.061 | 0.017 | 0.376 |

Table 2. Simulation results for bias, standard deviation and root mean squared error of $\hat{F}(t_{0.5})$ and $\hat{F}_C(t_{0.5})$

| | | | | $\hat{F}(t_{0.5})$ | | | $\hat{F}_C(t_{0.5})$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $c$ | $n$ | $q_T$ | bias | std | mse | bias | std | mse |
| 2.0 | 0.15 | 200 | 0.43 | -0.101 | 0.037 | 0.510 | -0.090 | 0.038 | 0.494 |
| 2.0 | 0.15 | 400 | 0.43 | -0.107 | 0.028 | 0.493 | -0.095 | 0.028 | 0.475 |
| 2.0 | 0.25 | 200 | 0.43 | -0.165 | 0.045 | 0.618 | -0.153 | 0.046 | 0.606 |
| 2.0 | 0.25 | 400 | 0.43 | -0.156 | 0.028 | 0.562 | -0.149 | 0.028 | 0.553 |
| 4.0 | 0.15 | 200 | 0.31 | -0.048 | 0.032 | 0.397 | -0.053 | 0.035 | 0.416 |
| 4.0 | 0.15 | 400 | 0.31 | -0.044 | 0.026 | 0.372 | -0.052 | 0.026 | 0.389 |
| 4.0 | 0.25 | 200 | 0.31 | -0.115 | 0.048 | 0.559 | -0.120 | 0.044 | 0.556 |
| 4.0 | 0.25 | 400 | 0.31 | -0.104 | 0.030 | 0.496 | -0.115 | 0.030 | 0.512 |
| 8.0 | 0.15 | 200 | 0.23 | -0.021 | 0.043 | 0.352 | -0.030 | 0.043 | 0.381 |
| 8.0 | 0.15 | 400 | 0.23 | -0.007 | 0.028 | 0.253 | -0.028 | 0.028 | 0.335 |
| 8.0 | 0.25 | 200 | 0.23 | -0.073 | 0.047 | 0.487 | -0.071 | 0.049 | 0.488 |
| 8.0 | 0.25 | 400 | 0.23 | -0.069 | 0.028 | 0.430 | -0.090 | 0.027 | 0.465 |

Table 3. Simulation results for bias, standard deviation and root mean squared error of $\hat{F}(t_{0.8})$ and $\hat{F}_C(t_{0.8})$

| | | | | $\hat{F}(t_{0.8})$ | | | $\hat{F}_C(t_{0.8})$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $c$ | $n$ | $q_T$ | bias | std | mse | bias | std | mse |
| 2.0 | 0.15 | 200 | 0.43 | 0.087 | 0.025 | 0.454 | -0.027 | 0.030 | 0.337 |
| 2.0 | 0.15 | 400 | 0.43 | 0.094 | 0.017 | 0.438 | -0.026 | 0.019 | 0.300 |
| 2.0 | 0.25 | 200 | 0.43 | 0.109 | 0.033 | 0.518 | -0.034 | 0.032 | 0.363 |
| 2.0 | 0.25 | 400 | 0.43 | 0.100 | 0.023 | 0.480 | -0.036 | 0.024 | 0.345 |
| 4.0 | 0.15 | 200 | 0.31 | 0.053 | 0.026 | 0.391 | -0.000 | 0.029 | 0.183 |
| 4.0 | 0.15 | 400 | 0.31 | 0.054 | 0.020 | 0.375 | -0.008 | 0.020 | 0.231 |
| 4.0 | 0.25 | 200 | 0.31 | 0.061 | 0.037 | 0.440 | 0.007 | 0.037 | 0.278 |
| 4.0 | 0.25 | 400 | 0.31 | 0.061 | 0.021 | 0.393 | -0.000 | 0.020 | 0.154 |
| 8.0 | 0.15 | 200 | 0.23 | 0.003 | 0.033 | 0.239 | 0.024 | 0.027 | 0.317 |
| 8.0 | 0.15 | 400 | 0.23 | 0.000 | 0.021 | 0.160 | 0.022 | 0.018 | 0.280 |
| 8.0 | 0.25 | 200 | 0.23 | -0.016 | 0.035 | 0.313 | 0.070 | 0.049 | 0.486 |
| 8.0 | 0.25 | 400 | 0.23 | -0.009 | 0.028 | 0.262 | 0.038 | 0.025 | 0.352 |

# Chapter 4

# Applications

For purpose of illustration, we apply the proposed method the CDC AIDS Blood Transfusion Data described in Example 2. To introduce interval censoring, we generate a uniform random variable $X$. If $X \leq 0.5$ then $E = \tau_D - 3.5$ and $R = \tau_D + 5.5$. If $X > 0.5$ then $E = \tau_D - 5.5$ and $R = \tau_D + 3.5$. The results of the estimators $\hat{F}(t)$ and $\hat{F}_C(t) = [\hat{F}(t) - \hat{F}(t_{0.1})]/[\hat{F}(t) - \hat{F}(t_{0.1})]$ are calculated using the constructed interval censored data. For purpose of comparison we also obtain the estimators of $F(t)$ and $F_C(t)$ using the exact observations, denoted by $\hat{F}_E$ and $\hat{F}_{EC}$, respectively. Figures 2 and 3 show the results of the estimators $(\hat{F}, \hat{F}_E)$ and $(\hat{F}_C, \hat{F}_{EC})$, respectively. Figures 2 and 3 indicate that except for the early times, both $\hat{F}$ and $\hat{F}_C$ are consistently smaller than $\hat{F}_E$ and $\hat{F}_{CE}$, respectively. The difference between $\hat{F}_C$ and $\hat{F}_{EC}$ is smaller than that between $\hat{F}$ and $\hat{F}_E$.

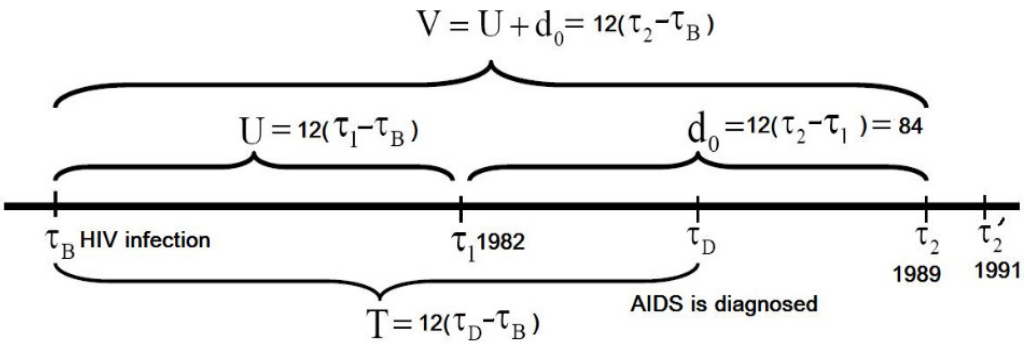


Figure 2. Schematic depiction of doubly truncated data

# Chapter 5

# Concluding Remarks

In this article, we have proposed a nonparametric estimate of the distribution function of $T$ when data is subject to double truncation and interval censoring. Simulation results indicate that the proposed estimator performs adequately when truncation is not severe. Furthermore, under certain assumptions on support of $T$, $(U, V)$ and $(E, R)$, we show the consistency of the conditional NPMLE. Further research is required to obtain the asymptotic results of the unconditional NPMLE.

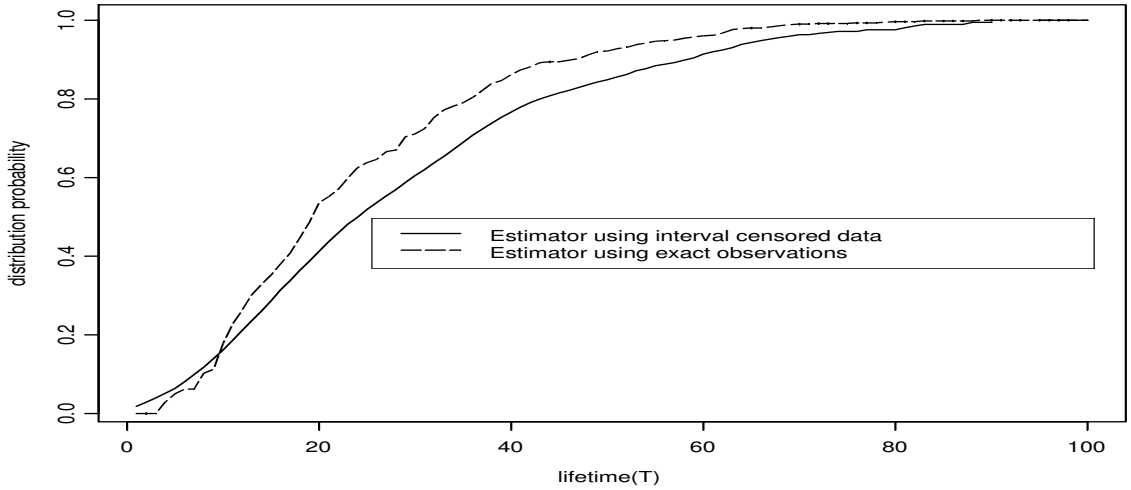


Figure 3. Plot of the results of the estimator $\hat{F}(t)$ and $\hat{F}_E(t)$

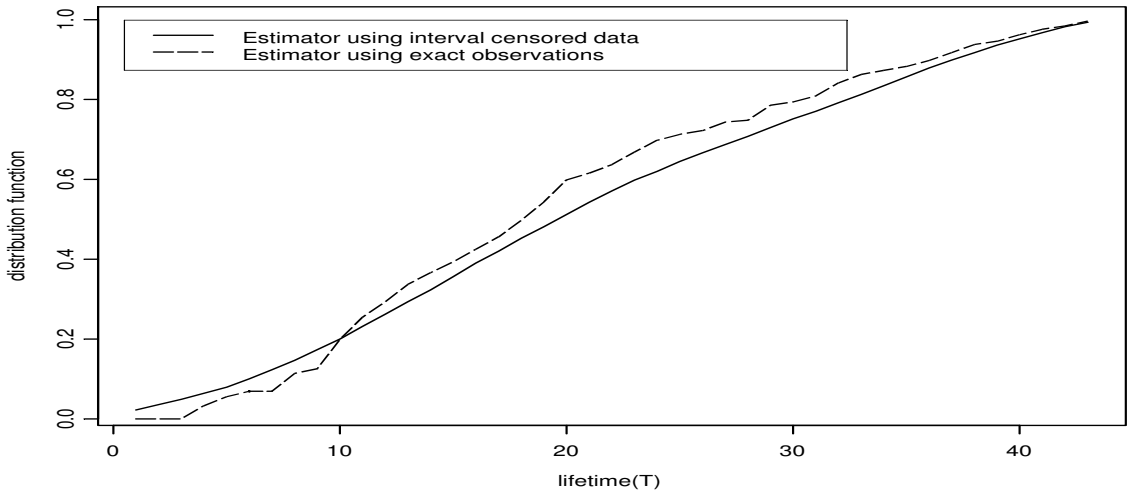


Figure 4. Plot of the results of the estimator $\hat{F}_C(t)$ and $\hat{F}_{EC}(t)$

# References


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

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.

Huang, J. and Wellner J. A. (1995). Efficient estimation for the proportional hazards model with case 2 interval censoring. Department of Statistics, University of Washington, Tech. Rept.

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.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457-481.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monograph National Royal Astronomical Society* **155**, 95-118.

Pan, W., Chappell, R. and Kosorok, M. R. (1998). On consistency of the monotone MLE of

survival for left truncated and interval-censored data. *Statistics & Probability Letters.* **38**, 49-57.

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

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

Shen, P.-S., 2008. Nonparametric analysis of doubly truncated data. *Ann. Inst. Stat. Math*, DOI10.1007/s10463-008-0912.2.

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

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

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

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

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

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

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

Woodroofe, M., 1985. Estimating a distribution function with truncated data. *The Annals of Statistics*, **13** 163-177.

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.