7th International Symposium on Imprecise Probability: Theories and Applications, Innsbruck, Austria, 2011

Bounds for Self-consistent CDF Estimators for Univariate and Multivariate Censored Data

Xuecheng Liu
Research Unit on Children's Psychosocial Maladjustment, University of Montreal, Canada
xuecheng.liu@umontreal.ca

Alain C. Vandal
Faculty of Health and Environmental Sciences, AUT University;
Centre for Clinical Research and effective practice, New Zealand
alain.vandal@aut.ac.nz

Abstract

In this paper, lower bounds and upper bounds are given for the mass assigned to a set of maximal cliques in self-consistent estimates of CDF NPMLEs for multivariate (including univariate) interval censored data under the assumption that the censoring mechanism is ignorable for the purpose of likelihood inference. The bounds are applied to give upper bounds of the diameter and size of the polytope of CDF NPMLEs for multivariate censored data.

Keywords. Interval censoring, maximal clique, clique matrix, self-consistent estimator, bounds, NPMLE, mixture nonuniqueness

1 Introduction

Survival analysis is the statistical analysis of event times, assumed nonnegative. It must account for, and is largely characterized by, censoring. Censoring is a type of coarsening of the data whereby an event time is only known up to an interval. While *right-censored data* consist of exactly observed times and intervals unbounded on the right, collections of positive values and of bounded and (right-) unbounded intervals on the nonnegative half-line are known as *interval censored data*. Right-censoring will occur in studies where follow-up is limited by design at a deterministic or random time. Interval censored data will typically arise in medical longitudinal studies, where patients can be assessed for a condition continuously, or at regular or irregular intervals.

The first task to undertake given interval censored data is often to estimate the underlying cumulative distribution function (CDF) F or equivalently the survival function $S = 1 - F$. In many instances, a nonparametric approach will be preferred to the constraining assumption of a parametric form for the CDF. In such situations the nonparametric maximum likelihood estimator (NPMLE) of the CDF will be the estimator of choice in the univariate case (Peto [18], Turnbull [22]), even when smoothing estimators are sought (Braun, Duchesne & Stafford [3]).

Event times can sometimes be stochastically associated, for instance through clustering. It is then useful to treat them as multivariate. Multivariate interval censored data are geometrically represented as the Cartesian product of the marginal event times or intervals that enter in a given observation.

Computing the CDF NPMLE can be a complex endeavor. Generally this computation can be carried out in two phases: in the first the effective support of the NPMLE is determined (Gentleman & Vandal [9], Bogaerts & Lesaffre [2], Maathuis [16]). This effective support consists in the *real representations (RR)* of the *maximal cliques* of the data, concepts to be defined in Section 2; for now it suffices to describe an RR as a generalized, possibly degenerate, hypercube in $\mathbb{R}_0^{+d}$, ($\mathbb{R}_0^+ = [0, +\infty)$, $d = 1, 2, ...$), with edges parallel to the axes. In the second phase a nonparametric likelihood with the CDF as argument is maximized; the maximizer assigns a probability mass to each RR (Wang [26]).

The probability vector obtained thus completely characterizes the CDF NPMLE. It is worth noting that this probability vector is always unique with univariate data (Vandal [23]). Arbitrary mass placement within an RR does not however affect the nonparametric likelihood, a situation to which we refer as *R-nonuniqueness* (Gentleman & Vandal [10]). With multivariate data, the probability vector itself may not be unique, a somewhat more serious situation we label *M-nonuniqueness*.

In this paper, we are interested in obtaining lower and upper bounds of the total CDF NPMLE mass assigned to an RR or a set of RRs of maximal cliques *without* conducting NMPL estimation. This is done by considering bounds on a class of more general estimators, namely self-consistent estimators (SCE), to

which NPMLEs belong.

These bounds can be obtained much more quickly than the probability vector that maximizes the likelihood (whether unique or not). There are good reasons for providing such bounds. First, even when one NPMLE vector is available, M-nonuniqueness will prevent us from deducing bounds for the probability mass on a *collection* of RRs. Second, reliable lower and upper bounds may enable us to select good starting probability vectors for NPML estimation: currently all algorithms used in for NPML estimation with general interval censored data are iterative. Third, there are self-contained applications of the bounds; in Section 5, we use them to provide upper bounds for the diameter and size of the polytope of NPMLEs.

The present paper focuses on nonparametric (and non-smoothed) maximum likelihood estimation. In that respect it differs from works such as those of Ferson et al. [7], whose statistical focus lies in parametric analysis with some forays in smoothing estimators. It also differs from the works such as that of Manski [17], that focus on the consequences of unobservability. This paper can be thought of as an inferential addition to the "catalogue" of techniques for symbolic data analysis, described in Billard & Diday [1].

We will assume in the sequel that the true CDF and the CDF NPMLE have support in $\mathbb{R}_0^{+d}$. We will also assume that the censoring mechanism is ignorable in the sense of Heitjian & Rubin [12], which implies in particular that likelihood-based inference relying on the data can be performed without reference to the censoring mechanism. A sufficient condition for ignorability of the censoring mechanism is for the underlying inspection process to be independent of the event times.

The rest of the paper is divided into 4 sections. In Section 2, we provide some necessary concepts and notation. In Section 3, we provide SCE bounds for any given collection of maximal clique RRs. In Section 4, we consider two special cases: one concerns the bounds on the SCE mass of a single maximal clique; the other the bounds on the SCEs given univariate censored data. In Section 5, we apply SCE bounds to give upper bounds of the diameter and size of the polytope of CDF NPMLEs for multivariate censored data.

2 Preliminaries and Notation

We provide some concepts and notation used in subsequent sections.

Let $R_1, ..., R_n$ be the n observations of an interval censored data set in $\mathbb{R}_0^{+d}$. Throughout this paper, we *always* assume that the censoring mechanism is ignorable in the sense of Heitjian & Rubin [12], which implies in particular that likelihood-based inference relying on the data can be performed without reference to the censoring mechanism. A sufficient condition for ignorability of the censoring mechanism is for the underlying inspection process to be independent of the event times. For any CDF F, the likelihood of F given the data is

$$L(F) = \prod_{i=1}^{n} P_F(R_i). \tag{1}$$

2.1 Intersection Graph, Maximal Clique, Clique Matrix, Real Representation

We can form the *intersection graph* of the data set in the following way: each observation corresponds to a vertex and two vertices are connected if and only if their corresponding observations intersect. A *clique* is a subset of vertices such that every pair are connected. A clique is called *maximal* if it is not a proper subset of another clique. The clique structure can be represented by the *clique matrix*, which is a 0/1 matrix, each row corresponding to a maximal clique and each column corresponding to an observation. An entry in the clique matrix is 1 if and only if the corresponding observation (i.e., vertex) belongs to the corresponding maximal clique. The clique matrix is unique up to permutations of rows and columns. In addition, each maximal clique has a real representation (RR), namely, the intersection of all its observations. The following is an illustrative example.

Example 2.1 *Let R_i, $i = 1, ..., 7$, be bivariate censored data as shown on Figure 1. Their intersection graph* [1] *is displayed in Figure 2. There are 4 maximal cliques M_1, M_2, M_3 and M_4:*

$$M_1 = \{R_1,\ R_2,\ R_4\}, \qquad M_2 = \{R_3,\ R_4,\ R_7\},$$
$$M_3 = \{R_4,\ R_5,\ R_6\} \quad \textit{and} \quad M_4 = \{R_4,\ R_6,\ R_7\}.$$

Their corresponding maximal intersections (i.e., real representations of maximal cliques) are shaded in Figure 1. The clique matrix of these data is given in Table 1.

2.2 NPML and Self-consistent Estimators of the CDF

The importance of maximal cliques lies in two facts: the possible support of NPMLE is limited to the RRs

[1] Note that each R_i intersects itself and hence corresponds to a loop in the intersection graph. The loops are ignored in Figure 1.

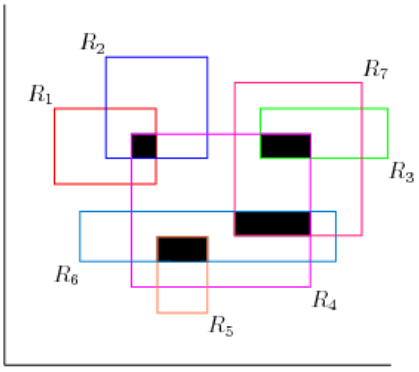

Figure 1: An example of bivariate interval censored data set

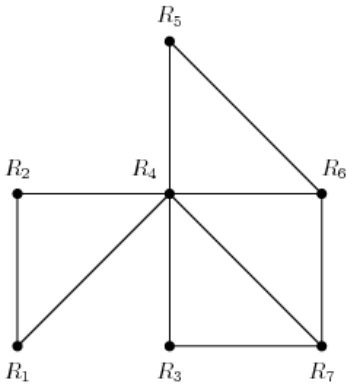

Figure 2: The intersection graph for the data in Figure 1

of maximal cliques; and the clique matrix is sufficient for the probability vector corresponding to the NPMLE. For a detailed discussion of the first fact, see Peto [18] and Turnbull [22]. For the second, refer to Gentleman & Vandal [10]. In the multivariate case, maximal cliques are most efficiently identified using the HeightMap algorithm of Maathuis [16] and the marked iso-graph algorithm (Liu [15]).

Suppose we have m maximal cliques $M_1, ..., M_m$, which are assigned masses $p_1, ..., p_m$ respectively. The likelihood (1) can then be redefined as a function of $\mathbf{p}$:

$$L(\mathbf{p}) = \prod_{j=1}^{n} \sum_{i=1}^{m} a_{ij} p_i, \tag{2}$$

where a_{ij}s valued in $\{0, 1\}$ are the entries of the clique matrix $\mathbf{A}_{m\times n}$. (That is, $a_{ij} = 1$ if and only if the observation R_j is in the maximal clique M_i.) The NPMLE corresponds to a probability vector $\mathbf{p} = [p_1, ..., p_m]'$. An NPMLE of the CDF will be constant except for increases of sizes p_i concentrated on the on the real representations of the maximal cliques.

	R_1	R_2	R_3	R_4	R_5	R_6	R_7
M_1	1	1	0	1	0	0	0
M_2	0	0	1	1	0	0	1
M_3	0	0	0	1	1	1	0
M_4	0	0	0	1	0	1	1

Table 1: Clique matrix of the data in Example 2.1

The precise placement of the mass within the real representations does not affect the likelihood, a situation to which we refer as R-nonuniqueness.

An important feature of a CDF NPMLE under censored data is that it must satisfy the self-consistency condition (Turnbull [22]). There are several equivalent definitions of self-consistency of estimators in the literature. We use the following, which precisely identifies fixed points of the EM algorithm:

Definition 2.2 *Let $\mathbf{A}_{m\times n}$ be the clique matrix for the multivariate censored data. A probability vector $\tilde{\mathbf{p}}$ is a self-consistent estimate if and only if*

$$n\tilde{\mathbf{p}} = \mathbf{D}_{\tilde{\mathbf{p}}}\mathbf{A}(\mathbf{A}'\tilde{\mathbf{p}})^{-\mathbf{I}}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix of order m, and

- $\mathbf{D}_{\mathbf{x}}$ denotes the diagonal matrix with diagonal $\mathbf{x}$;
- For any column vector $\mathbf{a}_{m\times 1} := [a_1, ..., a_m]'$, $a_i \neq 0$, $\mathbf{a}^{-\mathbf{I}}$ is the column vector whose $i-$th element is $1/a_i, i = 1, 2, ..., m$. [2]

The product-limit estimator for univariate right-censored data, first proposed by Kaplan & Meier [13], was later shown by Efron [6] to be self-consistent. Turnbull [21, 22] then used self-consistency as the basis for an estimation algorithm, later shown in Dempster, Laird & Rubin [5] to be a particular application of the EM algorithm. It is now a well recognized fact (Groeneboom & Wellner [11], Gentleman & Geyer [8], Wellner & Zhan [27]) that in general there exist several distinct values of $\tilde{\mathbf{p}}$ which are self-consistent but do not maximize the likelihood. In order to be the NPMLE, a self-consistent estimate must also satisfy the Kuhn-Tucker conditions listed in Gentleman & Geyer [8].

2.3 Further Notation

Let $\mathcal{C}$ be a set of maximal cliques of a multivariate censored data (MCD) set with n observations. Throughout this chapter, we use $\tilde{\mathbf{p}}$ to denote a self-consistent

[2] The notation $\mathbf{a}^{-\mathbf{I}}$ is a special case of Hadamard exponentiation. For more detailed information, see Gentleman & Vandal [9].

estimate. For such an estimate, define $\tilde{\mathbf{p}}_{\mathcal{C}}$ to be the total mass assigned to $\mathcal{C}$.

Let $n^+(\mathcal{C})$ and $n^-(\mathcal{C})$ be the numbers of observations in $\underset{C\in\mathcal{C}}{\cup} C$ and only in $\underset{C\in\mathcal{C}}{\cup} C$ respectively. Equivalently, we may interpret $n^+(\mathcal{C})$ as the number of observations covering some maximal clique RRs in $\mathcal{C}$ and $n^-(\mathcal{C})$ as the number of the observations covering only some maximal clique RRs in $\mathcal{C}$. Formally,

$$n^+(\mathcal{C}) := \Big|\bigcup_{C\in\mathcal{C}} C\Big|$$

and

$$n^-(\mathcal{C}) := \Big|\bigcup_{C_1\in\mathcal{C}} C_1 \setminus \bigcup_{C_2\notin\mathcal{C}} C_2\Big|.$$

We have

$$n^-(\mathcal{C}) = n - n^+(\mathcal{C}^c)$$

where $\mathcal{C}^c$ is the complement of $\mathcal{C}$ with respect to the set of all maximal cliques.

3 Bounds on Self-consistent CDF Estimates for MCD: General Case

3.1 Main Result

The main result of this section is the following theorem.

Theorem 3.1 *Let $\mathcal{C}$ be a set of maximal cliques of an MCD set with n observations, there holds*

$$\frac{n^-(\mathcal{C})}{n} \leqslant \tilde{\mathbf{p}}_{\mathcal{C}} \leqslant \frac{n^+(\mathcal{C})}{n}. \tag{4}$$

Proof. First, we prove the right-hand side of (4), that is

$$\tilde{\mathbf{p}}_{\mathcal{C}} \leqslant \frac{n^+(\mathcal{C})}{n}. \tag{5}$$

Without any loss of generality, we assume that in the clique matrix $\mathbf{A}$, the first $|\mathcal{C}|$ rows correspond to maximal cliques in $\mathcal{C}$ and first $n^+(\mathcal{C})$ columns correspond to observations in $\underset{C\in\mathcal{C}}{\cup} C$. Therefore, $\mathbf{A}$ is of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where the size of $\mathbf{A}_{11}$ is $|\mathcal{C}|$ by $|n^+(\mathcal{C})|$ and $\mathbf{O}$ denotes a matrix whose entries are all 0.

Rewrite $\tilde{\mathbf{p}}$ as $\tilde{\mathbf{p}} = \begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{bmatrix}$, where $\tilde{\mathbf{p}}_1 \in \mathbb{R}_+^{|\mathcal{C}|}$ and $\tilde{\mathbf{p}}_2 \in \mathbb{R}_+^{m-|\mathcal{C}|}$. Then $\tilde{\mathbf{p}}_{\mathcal{C}} = \sum_{i=1}^{|\mathcal{C}|} \tilde{p}_i$. Also let $\mathbf{I}$, $\mathbf{I}_1$ and $\mathbf{I}_2$ be the identity matrices of orders m, $|\mathcal{C}|$ and $m - |\mathcal{C}|$ respectively. The self-consistency condition on $\tilde{\mathbf{p}}$ becomes

$$\begin{aligned}
\begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{bmatrix} &= \frac{1}{n}\begin{bmatrix} \mathbf{D}_{\tilde{\mathbf{p}}_1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{\tilde{\mathbf{p}}_2} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\
&\quad \left(\begin{bmatrix} \mathbf{A}'_{11} & \mathbf{A}'_{21} \\ \mathbf{O} & \mathbf{A}'_{22} \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{bmatrix}\right)^{-\mathbf{I}} \\
&= \frac{1}{n}\begin{bmatrix} \mathbf{D}_{\tilde{\mathbf{p}}_1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{\tilde{\mathbf{p}}_2} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\
&\quad \begin{bmatrix} \mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2 \\ \mathbf{A}'_{22}\tilde{\mathbf{p}}_2 \end{bmatrix}^{-\mathbf{I}} \\
&= \frac{1}{n}\begin{bmatrix} \mathbf{D}_{\tilde{\mathbf{p}}_1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{\tilde{\mathbf{p}}_2} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\
&\quad \begin{bmatrix} (\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\ (\mathbf{A}'_{22}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_2} \end{bmatrix},
\end{aligned}$$

which implies that

$$\tilde{\mathbf{p}}_1 = \frac{1}{n}\mathbf{D}_{\tilde{\mathbf{p}}_1}\mathbf{A}_{11}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1}.$$

Hence, by letting $\mathbf{e}$ be the vector $[1]_{|\mathcal{C}|\times 1}$,

$$\begin{aligned}
\sum_{i=1}^{|\mathcal{C}|}\tilde{p}_i &= \mathbf{e}'\tilde{\mathbf{p}}_1 \\
&= \frac{1}{n}\mathbf{e}'\mathbf{D}_{\tilde{\mathbf{p}}_1}\mathbf{A}_{11}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\
&= \frac{1}{n}\tilde{\mathbf{p}}'_1\mathbf{A}_{11}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\
&= \frac{1}{n}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1)'(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1}.
\end{aligned}$$

Since $\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 \geqslant \mathbf{0}$ and $\mathbf{A}'_{21}\tilde{\mathbf{p}}_2 \geqslant \mathbf{0}$, we have

$$\begin{aligned}
\sum_{i=1}^{|\mathcal{C}|}\tilde{p}_i &\leqslant \frac{1}{n}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)'(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\
&= \frac{1}{n} \times n^+(\mathcal{C})
\end{aligned}$$

and (5) is proved.

Now we show the left part of (4). Denoting by $\mathcal{C}^c$ the complement of the set $\mathcal{C}$ of maximal cliques, we obtain from (5)

$$\tilde{\mathbf{p}}_{\mathcal{C}^c} \leqslant \frac{n^+(\mathcal{C}^c)}{n}$$

and therefore

$$\tilde{\mathbf{p}}_{\mathcal{C}} = 1 - \tilde{\mathbf{p}}_{\mathcal{C}^c} \geqslant 1 - \frac{n^+(\mathcal{C}^c)}{n} = \frac{n - n^+(\mathcal{C}^c)}{n} = \frac{n^-(\mathcal{C})}{n}.$$

The proof is complete. □

Note that, in the proof of Theorem 3.1, since $(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} > \mathbf{0}$ (without which the notation $(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 +$

$\mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1}$ does not make sense), the equality in (5) is valid (that is, $\tilde{\mathbf{p}}_{\mathcal{C}}$ reaches its upper bound in (5)) if and only if $\mathbf{A}'_{21}\tilde{\mathbf{p}}_2 = 0$. The latter condition is equivalent to the following statement: if the r^{th} entry in $\tilde{\mathbf{p}}_2$ is positive, then the r^{th} row of $\mathbf{A}_{21}$ is a zero row-vector.

Specifically, when $\mathbf{A}_{21}$ is the null matrix, $\tilde{\mathbf{p}}_{\mathcal{C}}$ reaches its upper bound in (5). In this case, the observations $\mathcal{R}$ corresponding to $\mathbf{A}$ can be divided into two groups: the observations which are only in $\mathcal{C}$ and observations only in $\mathcal{C}^c$. A similar argument is applicable to the left-hand side of (4). We can therefore conclude that (4) cannot be improved for any data set.

The lower and upper bounds described by Theorem 3.1 correspond to belief and plausibility measures in Dempster-Shafer Theory ([4, DST]). These measures are obtained from a basic assignment induced by equiprobability on the original data; this basic assignment is normalized to the set of maximal cliques rather than the power set of the data. To our knowledge this is the first time a relationship is established (via self-consistency) between Dempster-Shafer theory and non-smoothing/non-penalized nonparametric likelihood estimation.

3.2 Two Examples

Example 3.2 *Consider the data depicted in Figure 3: Applying (4) to the data, we obtain the bounds shown in Table 3. The "True region" heading indicates the bounds on the total mass of $\hat{\mathbf{p}}_{\mathcal{C}}$ implied by the M-nonuniqueness of the NPMLE. Indeed, the two end-points of the true region are the lower and upper probabilities defined by the NPMLE probability vectors.*

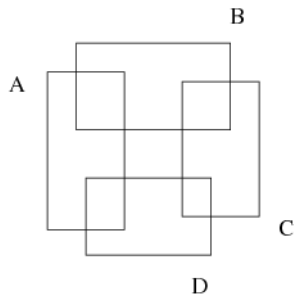

Figure 3: An example bivariate censored data set

Example 3.3 *Consider Pruitt's data (Pruitt [19]) depicted in Figure 4. Applying (4) to the data, bounds of $\tilde{\mathbf{p}}_{\mathcal{C}}$ for some given subsets $\mathcal{C}$ of maximal cliques are given in Table 5.*

	A	B	C	D
M_1	1	1	0	0
M_2	0	1	1	0
M_3	0	0	1	1
M_4	1	0	0	1

Table 2: The clique matrix for the data on Figure 3

$\mathcal{C}$	Lower bound	Upper bound	True region
$\{\tilde{p}_1\}, \{\tilde{p}_2\}, \{\tilde{p}_3\}, \{\tilde{p}_4\}$	0	2/4	[0, 0.5]
$\{\tilde{p}_1{+}\tilde{p}_2\}, \{\tilde{p}_2{+}\tilde{p}_3\}$, $\{\tilde{p}_3{+}\tilde{p}_4\}, \{\tilde{p}_4{+}\tilde{p}_1\}$	1/4	3/4	[0.5, 0.5]
$\{\tilde{p}_1{+}\tilde{p}_3\}, \{\tilde{p}_2{+}\tilde{p}_4\}$	0/4	4/4	[0,1]
$\{\tilde{p}_i{+}\tilde{p}_j{+}\tilde{p}_k,$ $1 \leqslant i < j < k \leqslant 4\}$	2/4	4/4	[0.5, 1]

Table 3: Mass bounds on $\mathbf{p}_{\mathcal{C}}$ for the data set in Example 3.2

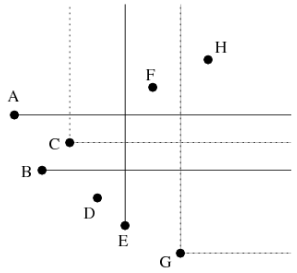

Figure 4: Pruitt's data set

	A	B	C	D	E	F	G	H
M_1	0	0	0	1	0	0	0	0
M_2	0	1	0	0	1	0	0	0
M_3	1	0	1	0	1	0	0	0
M_4	0	0	1	0	0	1	0	0
M_5	0	1	0	0	0	0	1	0
M_6	1	0	1	0	0	0	1	0
M_7	0	0	1	0	0	0	1	1

Table 4: The clique matrix for Pruitt's data set

$\mathcal{C}$	Lower bound	Upper bound	True region
$\{\tilde{p}_1\}$	1/8	1/8	[0.125, 0.125]
$\{\tilde{p}_2\}$	0/8	2/8	[0.095,0.191]
$\{\tilde{p}_5 + \tilde{p}_6\}$	0/8	4/8	[0.096, 0.096]
$\{\tilde{p}_2{+}\tilde{p}_3{+}\tilde{p}_5{+}\tilde{p}_6\}$	3/8	5/8	[0.457, 0.457]

Table 5: Mass bounds on $\mathbf{p}_{\mathcal{C}}$ for some $\mathcal{C}$'s for the data set in Example 3.3

3.3 Discussion

For uncensored data, the lower and upper bounds in (4) are always equal. Hence, (4) is an extension from uncensored to censored data.

M-nonuniqueness of NPMLEs for MCD can potentially create large differences between the lower and upper bounds in (4). From Examples 3.2 and 3.3, we notice that some intervals are wide and that we get no information at all in some cases. For instance, in Example 3.2, the lower and upper bounds for $\tilde{p}_1 + \tilde{p}_3$ are 0 and 1 respectively. Note, however, that tighter bounds on $\tilde{p}_1 + \tilde{p}_3$ are not available, since, the M-nonuniqueness interval of $\tilde{p}_1 + \tilde{p}_3$ is $[0, 1]$.

4 The Bounds in some Special Cases

4.1 Bounds on the SCE Mass of a Single Maximal Clique

In this section, we focus on the bounds for the mass assigned to one maximal clique by an SCE.

Theorem 4.1 *Let M_i be any maximal clique of an MCD set with n observations, there holds*

$$\frac{n^-(\{M_i\})}{n - n^+(\{M_i\}) + n^-(\{M_i\})} \leqslant \tilde{p}_i \leqslant \frac{n^+(\{M_i\})}{n}. \quad (6)$$

Note that, the lower bound in (6) improves the lower bound in (4) in Section 3, and the upper bounds in (6) and (4) are the same.

Proof of Theorem 4.1 . We only need to show the left-hand side of (6), that is

$$\tilde{p}_i \geqslant \frac{n^-(\{M_i\})}{n - n^+(\{M_i\}) + n^-(\{M_i\})}.$$

Denote by

$$\mathcal{J}_i := \{j; R_j \in M_i\}$$

the index set of $M_i \in \mathcal{M}$. So,

$$|\mathcal{J}_i| = n^+(\{M_i\}).$$

Also, denote

$$\tilde{\eta} := \mathbf{A}'\tilde{\mathbf{p}}.$$

Clearly, for every $i = 1, ..., m$ and all $j \in \mathcal{J}_i$,

$$\tilde{p}_i \leqslant \eta_j \leqslant 1. \quad (7)$$

Put

$$\mathcal{S}_i = \{j \in \mathcal{J}_i; R_j \text{ is only contained in } \mathcal{M}_i\}. \quad (8)$$

Then $|\mathcal{S}_i| = n^-(\{M_i\})$ and hence,

$$\begin{aligned} n &= \sum_{j\in\mathcal{J}_i} \frac{1}{\eta_j} \\ &= \frac{n^-(\{M_i\})}{\tilde{p}_i} + \sum_{j\in\mathcal{J}_i\setminus\mathcal{S}_i} \frac{1}{\tilde{\eta}_j} \quad (9) \\ &\geqslant \frac{n^-(\{M_i\})}{\tilde{p}_i} + \sum_{j\in\mathcal{J}_i\setminus\mathcal{S}_i} 1 \qquad [\text{from (7)}] \\ &= \frac{n^-(\{M_i\})}{\tilde{p}_i} + n^+(\{M_i\}) - n^-(\{M_i\}) \quad (10) \end{aligned}$$

whence the result follows.

Note that

$$n^+(\{M_i\}) = n^-(\{M_i\})$$

if and only if

$$n^-(\{M_i\})/(n - n^+(\{M_i\}) + n^+(\{M_i\}) = n^+(\{M_i\})/n.$$

□

4.2 Bounds on Self-consistent Estimates for Univariate Censored Data

In this section, we give the form of (4) and (6) for univariate censored data based on the characteristic matrix notation introduced by Vandal [23]. For univariate censored data, we further improve the lower bound for one maximal clique in (6).

4.2.1 Characteristic Matrix for Univariate Data

The clique matrix of a univariate censored data set is equivalent to its characteristic matrix, defined as follows.

Definition 4.2 *Let $\mathbf{A} = [a_{ij}]_{m\times n}$ be the clique matrix for a univariate censored data set $\{R_1, ..., R_n\}$ with maximal cliques $M_1, ..., M_m$ and corresponding RRs $H_1, ..., H_m$, ordered in the natural way. For each pair $i, j \in \{1, ..., m\}$ with $i \leqslant j$, define $\chi_{i,j}$ to be the number of columns in $\mathbf{A}$ such that the sub-column of 1's starts at i and ends at j.*[3] *The following upper-right triangle matrix*

$$\boldsymbol{\chi} := \begin{bmatrix} \chi_{1,1} & \chi_{1,2} & \cdots & \chi_{1,m-1} & \chi_{1,m} \\ & \chi_{2,2} & \cdots & \chi_{2,m-1} & \chi_{2,m} \\ & & \ddots & \vdots & \vdots \\ & & & \chi_{m-1,m-1} & \chi_{m-1,m} \\ & & & & \chi_{m,m} \end{bmatrix}$$

is called the characteristic matrix of the data. [4]

[3] Recall that the clique matrix of univariate censored data has the consecutive-1's property.

[4] The lower-left triangle in characteristic matrix is left undefined.

Example 4.3 *The following is the clique matrix of a univariate censored data set:*

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The equivalent characteristic matrix is

$$\boldsymbol{\chi} = \begin{bmatrix} 1 & 2 & 1 & 0 \\ & 0 & 3 & 0 \\ & & 0 & 2 \\ & & & 2 \end{bmatrix}.$$

4.2.2 Bounds on SCEs for Univariate Censored Data

The inequalities (4) for univariate censored data can be expressed using the entries of the characteristic matrix. Let $\mathbf{A}_{m\times n}$ be the clique matrix for a univariate data set with rows ordered according to the natural order of the maximal cliques[5]. Let $\tilde{\mathbf{p}} = [\tilde{p}_i]_{m\times 1}$ be a self-consistent estimate based on $\mathbf{A}$. Also, in this section, we always assume that $\chi_{1,m} = 0$ in $\mathbf{A}$'s characteristic matrix $\boldsymbol{\chi}$, since $\chi_{1,m}$ corresponds universal observations and have no bearing on the CDF estimation. For any given $j \in \{1, ...m\}$, let $\mathcal{C} := \{M_1, ..., M_k\}$. We then have

$$n^-(\mathcal{C}) = \sum_{\substack{s\leqslant k \\ r \text{ free}}} \chi_{rs} = \sum_{s=1}^{k}\sum_{r=1}^{s} \chi_{rs},$$

and

$$n^+(\mathcal{C}) = \sum_{\substack{r\leqslant k \\ s \text{ free}}} \chi_{rs} = \sum_{r=1}^{k}\sum_{s=r}^{m} \chi_{rs}.$$

From (4), the bounds on $\sum_{i=1}^{k} \tilde{p}_i$ can be given as

Theorem 4.4

$$\frac{1}{n}\sum_{s=1}^{k}\sum_{r=1}^{s}\chi_{r,s} \leqslant \sum_{i=1}^{k}\tilde{p}_i \leqslant \frac{1}{n}\sum_{r=1}^{k}\sum_{s=r}^{m}\chi_{r,s} \qquad (11)$$

Corollary 4.5 *When $j > 1$,*

$$\sum_{i=j}^{k}\tilde{p}_i \geqslant \frac{1}{n}\left(\sum_{s=1}^{k}\sum_{r=1}^{s}\chi_{r,s} - \sum_{r=1}^{j-1}\sum_{s=r}^{m}\chi_{r,s}\right), \quad (12)$$

$$\sum_{i=j}^{k}\tilde{p}_i \leqslant \frac{1}{n}\left(\sum_{r=1}^{k}\sum_{s=r}^{m}\chi_{r,s} - \sum_{s=1}^{j-1}\sum_{r=1}^{s}\chi_{r,s}\right). \quad (13)$$

[5] That is, $H < H'$ if and only if $x < x'$ for all $x \in H$ and $x' \in H'$.

Proof. Apply (11) to $\sum_{i=j}^{k}\tilde{p}_i$ and $\sum_{i=1}^{j-1}\tilde{p}_i$ and subtract. □

When we focus on the bounds of mass for one maximal clique, it is not difficult to show that for $i = 1, ..., m$,

$$\begin{aligned} n^-(\{M_i\}) &= \chi_{i,i}, \\ n^+(\{M_i\}) &= \sum_{j_1\leqslant i\leqslant j_2} \chi_{j_1,j2} =: n_i. \end{aligned}$$

from Theorem 4.1, we have

Theorem 4.6

$$\frac{\chi_{i,i}}{n - n_i + \chi_{i,i}} \leqslant \tilde{p}_i \leqslant \frac{n_i}{n}. \qquad (14)$$

If $0 < \chi_{i,i} < n_i < n$ for some $i = 1, ..., m$, then we can further improve the lower bound on $\tilde{p}_i$ in (14). First, for $i = 1, ..., m$, introduce the following notation:

$$\begin{aligned} l_i &:= \min\{r;\ \chi_{r,s} > 0 \text{ and } r \leqslant i \leqslant s\}, \\ u_i &:= \max\{s;\ \chi_{r,s} > 0 \text{ and } r \leqslant i \leqslant s\}, \\ d_i &:= \sum_{r=1}^{u_i}\sum_{s=r}^{m}\chi_{r,s} - \sum_{s=1}^{l_i-1}\sum_{r=1}^{s}\chi_{r,s} \end{aligned}$$

(We adhere to the usual convention that a summation over an empty index set is 0.) Since d_i is always smaller than n, the lower bound on $\tilde{p}_i$ can be improved in the following theorem.

Theorem 4.7

$$\tilde{p}_i \geqslant \frac{\chi_{i,i}d_i}{n(d_i - n_i + \chi_{i,i})}. \qquad (15)$$

Proof. The proof is similar to that of Theorem 4.1, except that for every $i = 1, ..., m$ and all $j \in \mathcal{J}_i$, $\tilde{\eta}_j$ now satisfies that,

$$\begin{aligned} \tilde{\eta}_j &\leqslant \sum_{c=l_i}^{u_i}\tilde{p}_c \\ &\leqslant \frac{d_i}{n}. \quad \text{[from (13)]} \end{aligned}$$

Therefore,

$$\begin{aligned} n &= \sum_{j\in\mathcal{J}_i}\frac{1}{\eta_j} \\ &= \frac{\chi_{i,i}}{\tilde{p}_i} + \sum_{j\in\mathcal{J}_i\setminus\mathcal{S}_i}\frac{1}{\tilde{\eta}_j} \\ &\geqslant \frac{\chi_{i,i}}{\tilde{p}_i} + \frac{n_i - \chi_{i,i}}{\frac{d_i}{n}}, \end{aligned}$$

and (15) follows. □

Corollary 4.8

$$\tilde{p}_1 \geqslant \frac{\chi_{1,1}(n-\chi_{m,m})}{n(n-\chi_{m,m}-n_1+\chi_{1,1})}$$
$$\tilde{p}_m \geqslant \frac{\chi_{m,m}(n-\chi_{1,1})}{n(n-\chi_{m,m}-n_i+\chi_{m,m})}$$

and for $i = 2, ..., m-1$,

$$\tilde{p}_i \geqslant \frac{\chi_{i,i}(n-\min(\chi_{1,1},\chi_{m,m}))}{n(n-\min(\chi_{1,1},\chi_{m,m})-n_i+\chi_{i,i})}.$$

Proof. Proof is obtained from the facts that

$$d_1 \leqslant n-\chi_{m,m}, \quad d_m \leqslant n-\chi_{1,1},$$

and for every $i = 2, ..., m-1$,

$$d_i \leqslant n-\min(\chi_{1,1},\chi_{m,m}).$$

□

Example 4.9 *Consider a univariate data set* $\{R_1, ..., R_{12}\} = \{1, 2, [3, 5], [4, 7], [6, 10], [8, 12], 9, [11, \infty), 13, [14, \infty), 15, [16, \infty)\}$ *which are represented in Figure 5. (The vertical positions hold no special meaning.) The RRs of the data are represented at the lowest vertical position and labeled* $H_1, ..., H_9$. *The*

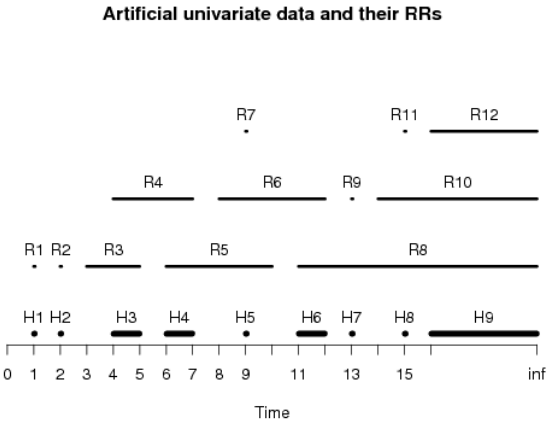

Figure 5: An artificial univariate data set

following is the clique matrix of this univariate censored data set:

$$\mathbf{A} = \begin{bmatrix} 1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&1&1&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&0&0&0&0&0 \\ 0&0&0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&1&1&0&0&0 \\ 0&0&0&0&0&0&0&1&0&1&1&0 \\ 0&0&0&0&0&0&1&1&0&1&0&1 \end{bmatrix}.$$

The equivalent characteristic matrix is

$$\boldsymbol{\chi} = \begin{bmatrix} 1&0&0&0&0&0&0&0&0 \\ &1&0&0&0&0&0&0&0 \\ &&1&1&0&0&0&0&0 \\ &&&0&1&0&0&0&0 \\ &&&&1&1&0&0&0 \\ &&&&&0&0&0&1 \\ &&&&&&1&0&1 \\ &&&&&&&1&0 \\ &&&&&&&&1 \end{bmatrix}.$$

The (unique) NPMLE probability vector is $\hat{\mathbf{p}} = [0.083, 0.083, 0.167, 0, 0.25, 0, 0.104, 0.156, 0.156]'$. *The CDF NPMLE is displayed in Figure 6.*

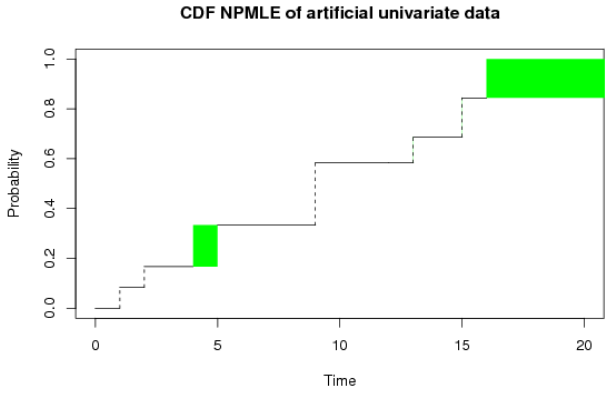

Figure 6: Example CDF NPMLE. Shaded boxes indicate areas of R-nonuniqueness, i.e. nonuniqueness related to arbitrariness of mass placement.

Applying Theorem 4.4, we can compare the NPMLE and the SCE lower and upper bounds as shown in Table 6.

$\mathcal{C}$	Lower bound	NPMLE	Upper bound
$\tilde{p}_1$	0.083	0.083	0.083
$\tilde{p}_1+\tilde{p}_2$	0.167	0.167	0.167
$\tilde{p}_1+\tilde{p}_2+\tilde{p}_3$	0.250	0.333	0.333
$\tilde{p}_1+\cdots+\tilde{p}_4$	0.333	0.333	0.417
$\tilde{p}_1+\cdots+\tilde{p}_5$	0.500	0.583	0.583
$\tilde{p}_1+\cdots+\tilde{p}_6$	0.583	0.583	0.667
$\tilde{p}_1+\cdots+\tilde{p}_7$	0.667	0.688	0.750
$\tilde{p}_1+\cdots+\tilde{p}_8$	0.750	0.844	0.917
$\tilde{p}_1+\cdots+\tilde{p}_9$	1.000	1.000	1.000

Table 6: NPMLE, lower and upper bounds comparison for a univariate data set

5 Application to M-Nonuniqueness

For MCD, the NPMLEs may display the aforementioned mixture or *M-nonuniqueness*, which occurs when different probability vectors have the same likelihood, that is, mass may be exchanged between maximal cliques without changing the likelihood. Gentleman & Vandal [9] proved that M-nonuniqueness cannot occur with univariate censored data but that it may arise with multivariate censored data. Liu [14, 15] and Vandal, Gentleman & Liu [24] discuss conditions for uniqueness of the NPMLEs and apply methods from convex optimization theory to show that the set of all NPMLEs is a polytope.

Suppose that the size of the clique matrix $\mathbf{A}$ for a MCD set is $m \times n$. When we have one CDF NPMLE $\hat{\mathbf{p}}$ for $\mathbf{A}$, the NPMLE's polytope $\mathcal{P}$ can be described by the following so called H-representation (Liu [14, 15], Vandal, Gentleman & Liu [24]):

$$\mathcal{P} = \{\mathbf{p} = [p_1, ...p_m]'; \mathbf{A}'\mathbf{p} = \mathbf{A}'\hat{\mathbf{p}}, \mathbf{p} \geqslant \mathbf{0} \text{ and } \mathbf{e}'\mathbf{p} = 1\}.$$

We consider three descriptions of the NPMLE's polytope $\mathcal{P}$ designed to quantify the extent of M-nonuniqueness. The first description is the so-called V-representation of $\mathcal{P}$, that is, the list of all its vertices. The second is the *diameter* of $\mathcal{P}$, corresponding to the longest distance between two of its vertices. The third is the *size* of $\mathcal{P}$, defined as the longest projection on one of the m axes corresponding to the vector entries. The diameter and size of $\mathcal{P}$ have been considered in the study of CDF NPMLE M-nonuniqueness and asymptotic properties. For more detail, see Liu [14, 15] and Vandal, Gentleman & Liu [24].

From (6), upper bounds for the diameter and the size of $\mathcal{P}$ for a censored data set with clique matrix $\mathbf{A}_{m\times n}$ can be obtained respectively as

$$\begin{aligned} \text{dia}(\mathcal{P}) &\leqslant \left(\sum_{i=1}^{m}(U_i - L_i)^2\right)^{1/2}, \qquad (16) \\ \text{size}(\mathcal{P}) &\leqslant \max_{i=1,...,m}(U_i - L_i), \qquad (17) \end{aligned}$$

where for $i = 1, 2, ..., n$,

$$L_i := \frac{n^-(\{M_i\})}{n - n^+(\{M_i\}) + n^-(\{M_i\})}$$

and

$$U_i := \frac{n^+(\{M_i\})}{n}$$

are lower and upper bounds of $\tilde{p}_i$ given in (6). As an application of (16) and (17), consider a cyclical data set with order $2k$, $k = 2, 3, ...$, circulant clique matrix defined as follows:

$$\begin{bmatrix} 1 & & & & & & 1 \\ 1 & 1 & & & & & \\ & 1 & 1 & & & & \\ & & \ddots & \ddots & & & \\ & & & 1 & 1 & & \\ & & & & 1 & 1 & \end{bmatrix} \qquad (18)$$

where all unspecified entries in the matrix are 0.

Then the diameter and size of the corresponding NPMLE polytope are at most $\sqrt{2/k}$ and $\frac{1}{k}$ respectively. [6]

The last theorem provides a sufficient condition for asymptotic mixture uniqueness.

Theorem 5.1 *When*

$$\frac{1}{n}\max_{i=1,...m}(n^+(M_i) - n^-(M_i)) \xrightarrow{\text{a.s}} 0,$$

the M-nonuniqueness of the CDF NPMLE will disappear asymptotically, in the sense that,

$$\text{size}(\mathcal{P}) \xrightarrow{\text{a.s}} 0.$$

Proof. Since for every $i = 1, ..., m$,

$$n^+(\{M_i\}) \geqslant n^-(\{M_i\}),$$

then from (17),

$$\text{size}(\mathcal{P}) \leqslant \max_{i=1,...,m}\Big(\frac{n^+(\{M_i\}) - n^-(\{M_i\})}{n}\Big).$$

The conclusion follows. □

Acknowledgements. The authors wish to thank two anonymous reviewers for their useful comments and suggestions.

References

[1] Billard, L. & Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *J Amer Statist Assoc* **98**, 470–487.

[2] Bogaerts, K. & Lesaffre, E. (2004). A new, fast algorithm to find the regions of possible support for bivariate interval-censored data. *J Comp Graph Statist* **13**, 330–340.

[6] We can in fact check that the diameter and size are $\sqrt{2/k}$ and $\frac{1}{k}$ exactly.

[3] Braun, J., Duchesne, T. & Stafford J.E. (2005). Local likelihood density estimation for interval censored data. *Can J Statist* **33**, 39–60.

[4] Dempster, A.P. (1967). Upper and Lower Probability inferences based on a sample from a finite univariate population. *Biometrika* **54**, 325-339.

[5] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximal likelihood estimation from incomplete data via the EM algorithm (with discussion). *J Roy Statist Soc B* **39**, 1–38.

[6] Efron, B. (1967). The two-sample problem with censored data. *Proc. Fifth Berkeley Symp Math Statist Probab* **4**, 831–853.

[7] Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W. & Ginzburg, L. (2007). Experimental uncertainty estimation and statistics for data having interval uncertainty. Sandia National Laboratories Technical Report SAND2007-0939.

[8] Gentleman, R. & Geyer, C.J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* **81**, 618–623.

[9] Gentleman, R. & Vandal, A.C. (2001). Computational algorithms for censored data using intersection graphs. *J Comp Graph Statist* **10**, 403–421.

[10] Gentleman, R. & Vandal, A.C. (2002). Graph-theoretical aspects of bivariate censored data. *Can J Statist* **30**, 557–571.

[11] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

[12] Heitjian, D. F. & Rubin, D.B. (1991). Ignorability and coarse data. *Ann Statist* **19**, 2244–2253.

[13] Kaplan E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J Amer Statist Assoc* **53**, 457–481.

[14] Liu, X. (2002). *Nonparametric Maximum Likelihood Estimation of the Cumulative Distribution Function with Multivariate Interval Censored Data: Computation, Identifiability and Bounds*. M.Sc. Thesis, Department of Mathematics and Statistics, McGill University, Montréal.

[15] Liu, X. (2005). *Nonparametric Estimation with Censored Data: a iscrete Approach*. Ph.D. Thesis, Department of Mathematics and Statistics, McGill University, Montréal.

[16] Maathuis, M.H. (2005). Reduction algorithm for the NPMLE for the distribution function of bivariate interval censored data. *J Comp Graph Statist* **14**, 252–262.

[17] Manski, C.F. (2003). *Partial Identification of Probability Distribution*. Springer-Verlag:Berlin.

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

[19] Pruitt, R.C. (1993). Small sample comparison of six bivariate survival curve estimators. *J Statist Comp Simul* **45**, 147–167.

[20] Rakowski, U.K. (2007). Fundamentals of the Dempster-Shafer Theory and its application to system safety and reliability modelling. *Reliability: Theory & Applications* (special issue) **3-4**, 173–185.

[21] Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J Amer Statist Assoc* **69**, 169–173.

[22] Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J Roy Statist Soc B* **38**, 290–295.

[23] Vandal, A.C. (1999). *Order theory and nonparametric analysis for interval censored data*. Ph.D. Thesis, Department of Statistics, University of Auckland.

[24] Vandal, A.C., Gentleman, R. & Liu, X. (2005a). Some comments on the uniqueness of the CDF NPMLE for censored data. Technical report, Department of Mathematics and Statistics, McGill University.

[25] Vandal, A.C., Gentleman, R. & Liu, X. (2005b). Constrained estimation and likelihood intervals for censored data, *Can J Statist* **33**, 71–83.

[26] Wang, Y. (2008). Dimension-reduced nonparametric maximum likelihood computation for interval-censored data, *Comp Statist Data Anal* **52**, 2388–2402.

[27] Wellner, J.A. & Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric likelihood estimator from censored data. *J Amer Statist Assoc* **92**, 945–959.