Reduction tree to determine estimators from the multivariate interval censored data

M Fatekurohman1, T A Kusmayadi2 and D P Antika1
1Department of Mathematics, University of Jember, Indonesia
2Department of Mathematics, Universitas Sebelas Maret, Indonesia

Email: mfatekurohman.fmipa@unej.co.id

Abstract. In this paper methods for finding the nonparametric maximum likelihood estimator (NPMLE) of the distribution function of time to event data will be presented. The basic approach is to use graph theory (in particular reduction tree), to simplify the problem. The Multivariate interval censored data can be represented by its intersection graph. The idea behind the reduction tree approach is based on the fact that when censored data can be partitioned into several groups such that observations from different groups do not intersect, we may compute the NPMLE with the data in each group separately. Each group can be further simplified when some observations in the group intersect all other observations: with respect to the group, such observations have no incidence on the NPMLE computation. Once these observations are, in effect, removed from the group, the group may be further partitioned. The process can go on recursively, forming the reduction tree. In the three dimension, the support area intersection of beams, where $X$ abis the time of infection, $Y$ ordinat the time of onset and $Z$ the time of death. The results obtained by the leaves from the reduction tree are not unique and are maximum cliques.

1. Introduction

Estimation of a survival function is perhaps the first and most commonly required task in the analysis of failure time data. There can be many reasons or purposes for such a task. For example, an estimated survival function can be used to assess the validity of an assumption about a particular parametric model for the underlying survival variable of interest. Also, one may need to estimate survival functions to estimate certain survival probabilities, to graphically compare several different treatments, or to predict survival probabilities for patients. In the case where a parametric model can be reasonably assumed for the underlying survival function, the estimation problem is relatively easy, and the maximum likelihood approach, is commonly used for the problem.

In the case of right-censored failure time data, the nonparametric maximum likelihood estimator (NPMLE), the topic that has been discussed the most in the analysis of interval-censored failure time data, and the study of it goes back to the fifties. According to [1] and Eeden (1956) in [10], were the first to derive the NPMLE of a distribution function based on current status data, [9] and [11], investigated the estimation problem based on general or case II interval-censored data. For the problem, the former presented a Newton-Raphson algorithm and the latter developed the self-consistency algorithm, Kaplan-Meier estimator [7] & [6]. It is a product-limit estimator and has been extensively studied in the literature. Furthermore, its pointwise variance estimate is available and given by the well-known Greenwood’s formula [5]. For interval censored failure time data, unlike parametric inference, nonparametric inference is much more complicated than that for right-censored
data from both practical and theoretical points of views. In particular, the NPMLE of a survival function does not have a closed form in general and can only be determined using iterative algorithms. The main focus of this paper is determine estimators using tree graphs has been discussed among them, [8] determine estimator used tree reduction of univariate censored data, [3] determine estimator used rooted tree of bivariate censored data. This paper will be discussed looking for estimators in multivariate interval censored data using reduction tree.

2. Methods

Definition 2.1
A graph is a finite nonempty set $V(G)$ of objects called vertices (also called points or nodes), and a (possibly empty) set $E(G)$ of 2-element subsets of $V(G)$ called edges (or lines). The set $V(G)$ is called the vertex set of $G$ and $E(G)$ its edge set. Let $G$ be a graph and $\{u,v\}$ an edge of $G$, since $\{u,v\}$ is two element set, we may write $\{u,v\}$ instead of $\{u,v\}$. It is often more convenient to represent this edge by $uv$ or $vu$. If $e=uv$ is an edge of a graph $G$, then we say that $u$ and $v$ are adjacent in $G$, and that $e$ joins $u$ and $v$. The graph $G$ itself is connected if $u$ is connected to $v$ for every pair $u$, $v$ of vertices of $G$. [2]

Definition 2.2
A clique is a subset $C$ of $V$ such that every member $C$ is connected by an edge to every other member of $C$. A clique is a maximal if it is not a proper of any other clique.

In the context of censored data, a clique is a set of indices corresponding to individual whose regions (i.e. real representations) intersect, is the regions of intersection of its members. The maximal cliques denote are $M_i$, and we use $M = \{M_1, M_2, ..., M_m\}$, [8], to denote the set of maximal cliques, [4]. Maximal clique of an intersection graph on censored data plays a very important role in the MLE. The structure of maximal cliques intersection graph can be summarized into a clique matrix, a matrix of $0/1$, so that each row state a maximal clique and each column is for one observation. If the value $a_{ij} = 1$ then the observation corresponds maximal cliques. A clique matrix is defined as

$$
A = \begin{bmatrix}
    a_{11} & a_{12} & \cdots & a_{1n} \\
    a_{21} & a_{22} & \cdots & a_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{m1} & a_{m2} & \cdots & a_{mn} \\
\end{bmatrix}
$$

where $M_1, M_2, ..., M_m$ are the maximal cliques of the intersection graph data and $a_{ij} = 1$ ($R_i \in M$), for all $i=1,2, ..., m$ and $j=1,2, ..., n$ [8].

2.1 Principles of Data Reduction

The Matrix used is the result of beam intersection matrix that has been simplified into clique matrix. According to [8], if $\mathcal{R}$ is dividable into groups $\mathcal{R}_g$, $g = 1, 2, 3, ..., G$, with $G \geq 2$, then, to get NPMLE for each group $\mathcal{R}_g$ separately with group total mass $|\mathcal{R}_g| = |\mathcal{R}|$.

Case 1: If $\mathcal{R}$ is dividable then we let the groups for $\mathcal{R}$‘s, say, $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, ..., \mathcal{R}_k$ be the children of $\mathcal{R}$.

Case 2: If $\mathcal{R}$ has universal observation (\mathcal{R} is not dividable) then we let $\mathcal{R}^{(*)}$ be the only child of $\mathcal{R}$.

Case 3: If $\mathcal{R}$ neither contains universal observations nor is dividable then we let $\mathcal{R}$ be a leaf.

The recursive approach is used for each $\mathcal{R}_i$ (Case 1) and $\mathcal{R}^{(*)}$ to produce the reduction tree. The leaves of reduction tree are nondividable and free of universal observations. Moreover, the set of leaves of reduction tree is unique.

According to [8], an observation universal if it intersects all other observations. For data set $\mathcal{R}$ denote by $\mathcal{R}^{(*)}$ the data set $\mathcal{R}$ in which aU universal observations are removed. When data set $\mathcal{R}$ can be partitioned into $G$ groups $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, ..., \mathcal{R}_G$, such that observations from different groups do not intersect, we say that the data set is dividable.
The clique matrix $A_{mxn}$ of a dividable data set $R$ can be written in the form $A = \begin{bmatrix} A_1 \\ \vdots \\ A_G \end{bmatrix}$, where $[A_1]_{n1x1}, \ldots, [A_G]_{n GX m}$. are cliques matrix, for $R_1, R_2, \ldots, R_G$, and $m = m_1 + m_2 + \ldots + m_G$, and $n = n_1 + n_2 + \ldots + n_G$.

Suppose $p$ partitioned into groups such as $p = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_G \end{bmatrix}$, self consistent condition $n p = D p A (A' p)^{-1}$, is equivalent to $n p_g = D p_g A (A' p_g)^{-1}$, $g = 1, 2, 3, \ldots, G$. Therefore if $\hat{p} = \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_G \end{bmatrix}$ is the NPMLE for $A$, for each $g = 1, 2, 3, \ldots, G$, then $n e' \hat{p}_g = e' D \hat{p}_g A_g (A' \hat{p}_g)^{-1}$

$$= \hat{p}_g A_g (A' \hat{p}_g)^{-1}$$

$$= (A_g \hat{p}_g)'(A_g \hat{p}_g)^{-1}$$

$$= n_g.$$ 

Here the probability mass allocated to the group $g$ by NPMLE is $n_g/n$.

Recursive approach for each $R_j$ (case1) and $R^*$ produce a reduction tree, the leaves of the root of the tree cannot be divided and free from universal observation, because the set of leaves of a single tree is rooted. NPMLE mass divided into each leaf can be obtained in the formation of the tree rooted with the road, put on the root mass is 1. Suppose $R^*$ and $R^{(*)}$ is a collection of observations in a vertex in an arbitrary level in the tree roots, waste mass will not change between $R^*$ and $R^{(*)}$. When $R^*$ (or $R^{(*)}$) be shared, the roots of the sub-tree $R^*$ (or $R^{(*)}$) will be assigned to the mass proportional to the number of observations in each vertex, with equal mass total $R^*$ Reduction tree construction process, that total mass NPMLE always occupying ever single leaf. To get NPMLE given by the original data, focused only on the leaves of the tree rooted. Suppose $R_1, R_2, R_3, \ldots, R_L$ data set corresponding to the leaves of the tree rooted with mass $m_1, m_2, m_3, \ldots, m_L$, and if $\hat{p}_k$ NPMLE is a probability vector for a given set of data $R_k$ with $k = 1, 2, 3, \ldots, L$, the NPMLE obtained from the original data set is

$$\hat{p} = \begin{bmatrix} m_1 \hat{p}_1 \\ m_2 \hat{p}_2 \\ \vdots \\ m_L \hat{p}_L \end{bmatrix},$$

with $\hat{p}_k$ NPMLE is obtained from each of the original data set $R_k$, for $k = 1, 2, 3, \ldots, L$.

2.2 Principles of Reduction Trees

The above discussion leads to the establishment of the following two simplifying principles in the computation of the NPMLE.
1. Universal observations play no role in finding NPMLE and we can simply ignore them for estimation purposes. Thus we need only consider data sets in which all universal observations are removed.

2. If a data set is divisible into groups \(R_i, i=1,2,3,\ldots, |G|\), with \(|G| \geq 2\), then we can find the NPMLE for each group \(R_i\) separately with group total mass \(|R_i|/|R|\).

We can recursively apply the above two principles to form the reduction tree of the data, as described below. The tree's nodes are subsets of the original data set \(R\), while the root of the tree is set to be \(R\) itself.

3. Result

Searching for the NPMLE for censored data can be divided into two steps: the first is the determination of all maximal intersections which form the support of the NPMLE; the second is to compute the NPMLE vector assigning mass only to maximal intersections. In this section, we focus on the second step: determining the estimator of multivariate interval censored data.

3.1 Reduction Tree for Multivariate Censored Data

The reduction tree algorithm of multivariate interval-censored data is development of reduction tree algorithm by [4] and [2]. Case of univariate data, algorithm of [4], adjacency matrix obtained is a square matrix and can be used directly to calculate the estimator using the reduction tree algorithm.

a). The steps algorithm of [4], for univariate data as follows:

- Step 1. given adjacency matrix \(A_{\text{adj}}\) of intersect graph as root.
- Step 2. matrix \(A_{\text{uni}}\) contains universal observations,
  2a). If it contains universal observations, the probability mass is 0,
  2b). If it is not contains then the probability mass can be obtained from the rest of the matrix.
- Step 3. repeat step 2 for the rest matrix until all of the matrix ordo has been processed.

b). The steps reduction tree algorithm in bivariate censored data [2] as follows:

- Step 1. given adjacency matrix intersect graph.
- Step 2. simplified into clique matrix.
- Step 3. clique matrix is placed as root
- Step 4. clique matrix not contains universal observation and cannot dividable and the result is leaves that not unique.

Whereas for multivariate interval-censored data using beam matrix so that to calculate estimator, a simplification step must be needed from the adjacency matrix into the clique matrix form. This step is needed to facilitate the calculation process of obtaining estimators through reduction tree. The leaves obtained from the reduction tree algorithm are not unique and are maximum cliques. The steps of reduction tree algorithm in multivariate censored is the same as bivariate (step 1, 2, 4), but in step 3 the clique matrix is checked one by one, because each clique matrix obtained will produce an estimator and give a maximum value of different functions, the steps as follows:

- Step 1. given adjacency matrix intersect graph.
- Step 2. simplified into clique matrix.
- Step 3. try one by one from clique matrix as the root.
- Step 4. if the clique matrix has not universal observations and cannot be divided then leaves are non unique.

3.2 A Simple example

Given a simple example of six of data of patients, each patient has time to be infected, onset and death, can be representative in three dimension. It can be described in the three-dimensional space, where \(X\) denote the time of infection, \(Y\) denote the time of onset and \(Z\) the time of death. Every data can be expressed as \((x_{1i}, x_{2i}, y_{1i}, y_{2i}, z_{1i}, z_{2i}), \ i = 1, 2, \ldots, 6\). i.e. \(B_1 = (1, 3, 5, 2, 4, 6), B_2 = (2, 2, 4, 4, 6, 7), B_3 = (10, 12, 13, 15, 17, 20), B_4 = (5, 6, 8, 9, 13), B_5 = (4, 4, 6, 7, 10)\) and \(B_6 = (3, 7, 8, 6, 10, 15)\).
Figure 1. A simple example of six observations

from Figure 1, there are six observation beams $B_1, B_2, B_3, B_4, B_5$ and $B_6$. Table 1. Express the intersections $M_1 = \{B_1, B_2\}$, $M_2 = \{B_2, B_5\}$, and $M_4 = \{B_4, B_5, B_6\}$, maximal cliques, but $M_3 = \{B_3\}$ is not maximal. Real representations are exactly the inner beams at which the maximum likelihood estimation can place probability mass. Both Figure 1 and 2, reflect the fact that $B_4B_5 \in E$ while $B_4B_5 \notin E$. The incidence matrix for this example is given Table 2 and the clique graph is displayed in Figure 3.

Table 1. The intersection matrix of six observations

\[
\begin{bmatrix}
B_1 & B_2 & B_3 & B_4 & B_5 & B_6 \\
M_1 & 1 & 1 & 0 & 0 & 0 \\
M_2 & 0 & 1 & 0 & 0 & 1 \\
M_3 & 0 & 0 & 1 & 0 & 0 \\
M_4 & 0 & 0 & 0 & 1 & 1 \\
M_5 & 0 & 0 & 0 & 1 & 1 \\
M_6 & 0 & 0 & 0 & 1 & 1
\end{bmatrix}
\]
A graph of the data can be constructed by making each rectangle a vertex, and then joining by edges those vertices whose rectangles intersect (Figure 2) and Figure 3 vertex E’s adjacency set and it is simplicial of Figure 2. The adjacency sets of the other four vertices each consist of two points and these are not connected by an edge, [3].

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

3.3 Algorithm for Multivariate Interval Censored Data

Level 0. (Root).

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

We can see that $\mathcal{R}$ has no universal observations and can divided into $\mathcal{R}_1 = \{R_1\}$, $\mathcal{R}_2 = \{R_2\}$, and $\mathcal{R}_3 = \{R_3\}$, dan $\mathcal{R}_4 = \{R_4\}$, with masses $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{4}$.

Level 1. $\mathcal{R}_{1,1}, \mathcal{R}_{1,2}, \mathcal{R}_{3}$ and $\mathcal{R}_{1,4}$ are leaves.

Figure 2. Intersection graph of the beams

Figure 3. Clique graph of the beams
Illustration of $R^*$ dan $R^{(r)}$ are showed as reduction tree

![Reduction tree of the data set](attachment:reduction_tree.png)

Figure 4. Reduction tree of the data set

Leaves of reduction tree are $\{R_1\}$, $\{R_2\}$, $\{R_3\}$, and $\{R_4\}$ with each mass allocation $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{4}$, four leaves are maximal cliques and NPMLE that can be expressed with $R$ is $\hat{\mathbf{p}} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & 4 \end{pmatrix}$.

4. Conclusion

In this paper has been discussed about determination the estimator using reduction tree and the results are, the leaves obtained from the reduction tree are not unique and are maximum cliques. The main advantage of finding a reduction tree is the data multivariate NPMLE simplification of NPMLE computation. For multivariate data, the algorithm’s performance needs more detailed study for different censoring schemes.

Suggestions for further research, for application, we could estimate the “trivariate” distribution function of the time of HIV infection, the time of onset of AIDS and the time of death from AIDS. We could be left, right, interval and double censored.

Acknowledgement

The authors would like to thank the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia for supporting this research through Grant of Pasca Doctor 2018, Contract Number: 1258/UN25.3.1/LT/2018.

References

[1] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. 1955. An empirical distribution function for sampling with incomplete information. The Annals of Mathematical Statistics, 26, 641-647

[2] Chartrand, G. and Oellermann, O.R., 1993. Applied and Algorithmic Graph Theory, McGraw-Hill, Inc., New York

[3] Fatekurohman.M, Subanar and Danardono. 2014. An Algorithm of Nonparametric Maximum Likelihood Estimation for Bivariate Censored Data. 1st ISIM-MED 2014. Innovation and Technology for Mathematics and Mathematics Education” Department of Mathematics Education,Yogyakarta State University Yogyakarta, pp. SP21-SP26

[4] Gentleman, R. and Vandal, A.C. 2002. Graph Theoretical Aspects of Bivariate Censored Data. Can. J Statist 10, pp 557-571

[5] Greenwood M. 1926. The natural duration of cancer. Reports on Public Health and Medical Subjects, 33, 1-26.

[6]Kalbfleisch, J. D. and Prentice, R. L. 2002. The statistical analysis of failure time data. Second edition, John Wiley: New York.

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

[8] Liu.X. 2005. Nonparametric estimation with censored data: a discrete approach: Department of Mathematics and Statistics McGill University, Montreal

[9] Peto, R. 1973. Experimental survival curves for interval-censored data. Applied Statistics, 22, 86-
91

[10] Sun, J. 2006. *The statistical analysis of interval-censored failure time data*. New York: Springer.

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