Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 1 (2015), pp. 47-55

http://www.ripublication.com

Algorithm of Intersection Graph for Bivariate Censored Data

Mohamad Fatekurohman[1], Subanar[2] and Danardono[3]

[1]Department of Mathematics, Jember University
Ph.D. student, Department of Mathematics, Gadjah Mada University
mfatekurohman@gmail.com
[2,3]Department of Mathematics, Gadjah Mada University
{subanar, danardono}@ugm.ac.id

Abstract

Univariate or bivariate Censored data, left, right or interval censored data can be represented by its intersection graph. Determination of nonparametric maximum likelihood estimator for bivariate interval censored data consists of two parts. The first, involves the determination of the regions of possible support and the second the maximization of the likelihood. As first step in the estimation of the nonparametric maximum likelihood estimation (NPMLE), for bivariate interval censored data, the regions of possible support, i.e. the rectangles with nonzero mass, are calculated. In this paper we discuss a method for finding an intersection graph. As the results, an algorithm of intersection graph is obtained and its time complexity is quadratic. This algorithm is an alternative way to calculate the regions of possible support for bivariate interval censored data.

AMS Subject Classification: 68R10, 62N01

Key words and Phrases: bivariate censored data, intersection graph, NPLME, time complexity.

1. Introduction

Dengue fever is a mosquito-borne tropical disease caused by the dengue virus. After a person is bitten by an infective mosquito, the virus undergoes an incubation period of 3 to 14 days (average 4 to 7 days), after which the person

may experience acute onset of fever accompanied by a variety of nonspecific signs and symptoms. During this acute febrile period, which may be as short as 2 days and as long as 10 days, dengue viruses may circulate in the peripheral blood (Gubler, [5]). Related to bivariate data, a patient bitten by an infective mosquito the virus undergoes an incubation period is representated by x-axis and a patient infected with the virus is represented by y-axis.

Consider a survival study that involves n independent subjects from a homogeneous population with each subject giving rise to two failure times denoted T_{1i} and T_{2i}, $i = 1, ..., n$, (Sun, [8]). Let $F(t_1,t_2)=P(T_{1i} \leq t_1, T_{2i} \leq t_2)$ denote their joint cumulative distribution function and suppose that only interval censored failure time data are available. In particular, the observations are

$$U_i = (L_{1i},R_{1i}] \times (L_{2i},R_{2i}] \, , i = 1, \ldots , n \}$$

where $(L_{1i},R_{1i}]$ dan $(L_{2i},R_{2i}]$ represent the intervals to which T_{1i} and T_{2i}, respectively. The observation on each subject could be a point, line segment (which may be a half-line), or rectangle (which may be a quadrant). These possibilities correspond to the situations where both failure times are observed exactly, one failure time is observed exactly and the other is interval or right censored, or both failure times are interval or right censored, respectively. If one treats points as rectangles that are degenerate in both dimensions and line segments as rectangles that are degenerate in one dimension, then the observed data consist entirely of rectangles. That is, the observed data are a collection of n rectangles. As before, we use the convention that $(a, a]$ means the single point $\{a\}$.

According to Bogaerts and Lesaffre [1], there are two steps that must be done for finding nonparametric maximum likelihood estimator of bivariate interval censored data. The first determines the intersection area of rectangles and the second finds the maximum likelihood. There are some methods to find intersection graph and some researcher found that time complexity of the algorithm is difference [1], [3] and [7].

2. Graph Theory, Intersection Graph, Maximal Intersection, Maximal Cliques and Clique Matrix.

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 2 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, (Chartrand and Oellermann, [2]).

A clique is subset C of V such that every member of C is connected by an edge other member of C. A clique is said to be maximal if the clique is not a

proper subset of any other clique. In the context of censored data, a clique is a set of indices corresponding to individuals whose region(i.e real representation) intersect. The real representation of the clique is the region of intersection of its members, maximal cliques denote are M_i. We use $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$ to denote the set of maximal cliques, (Gentleman and Vandal, [4]). Maximal clique of an intersection graph on censored data plays a very important role in the MLE. The structure of maximal clique 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

$$\begin{array}{c} \quad R_1 \quad R_2 \quad . \quad . \quad R_n \\ \mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & . & . & a_{1n} \\ a_{21} & a_{22} & . & . & a_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{m1} & a_{m2} & . & . & a_{mn} \end{bmatrix} \begin{array}{c} M_1 \\ M_2 \\ . \\ . \\ M_m \end{array} \end{array}$$

where $M_1, M_2, \ldots, M_m$ are the *maximal cliques* of the data intersection graph and $a_{ij} = 1$ ($R_j \in M_i$), for all $i = 1,2,\ldots,m$ and $j = 1,2,\ldots,n$. $= 1$, (Liu, [6]).

According to Gentleman and Vandall [3], we now consider the application of this theory to censored data. Every observation belongs to at least one maximal clique. For every finite data set there are a fixed number m maximal cliques. Let H_i denote the real representation of the region defined of the members of M_j that is, $H_i \overset{def}{=} \bigcap_{j \in M_i} R_j$. $\mathcal{H}$ is the set of the regions corresponding to the maximal cliques of the intersection graphs, denote $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$, so that there is one-one correspondence between $M_i \leftrightarrow H_i$, $i = 1,2,3,\ldots, m$. Later we will show that the maximum likelihood estimate of F concentrates on (possibly a subset of) the H_i and hence on the maximal cliques. It is worth nothing that the maximal cliques involve only the set of indices of those individuals in the clique. The real representation of the cliques (i.e where it is on the line or the plane) is largely irrelevant to maximum likelihood.

A region H is an intersection of R_1, R_2,..., R_n maximal if H is nonempty intersection of R_i's and there is no intersection H_i, so H_i is a proper subset of H. Given censored data, for any real representation of the maximal clique is a maximal intersection and that any maximal intersections is representation of a maximal cliques. Maximal intersections and real representations of maximal cliques are thus identical, and no confusion can arise inrelating the set of maximal intersections to the set of maximal cliques, (Liu, [6]). Maximal clique is very important because the NPMLE estimates of F concentrated on maximum cliques, as well as on H_i.

Incidence matrix A is an n x n matrix with entries α_{ij}, if $\alpha_{ij} = 1$ if the

individual j is a maximal clique i and 0 for the others. When connectedness with interval graph, the incidence matrix refers to a matrix cliques.

3. Graph Structure in Censored Data

Assumed that bivariate event times are independently distributed according to the distribution $F(t_1,t_2)$. It is this distribution, or the corresponding survivor function that is of interest. The exact data $\{t_{ij},t_{2j}\}_{j=1}^{n}$, are not observed rather the data are censored according to some process. One assumption could be that there is an inspection time process Q_j (t_1,t_2) associated with each individual. Suppose given data set $\Re=\{R_1,R_2,R_3,\ldots,R_n\}$, where coordinat $(x_{lj},y_{lj},x_{uj},y_{uj})$ for R_j and $t_{1j}\in(x_{lj},x_{uj})$, $t_{2j}\in(y_{lj},y_{uj})$. An intersection graph formed from the set of data $\Re$, that is, each observation R_j corresponds to the vertex v_j. Two vertices v_j and v_k are connected by an edge if two rectangles intersect R_j and R_k, with $j\neq k$. The two rectangles R_i and Rj said intersection if the least one point (coordinates) of R_i is contained within the rectangle R_j and otherwise.

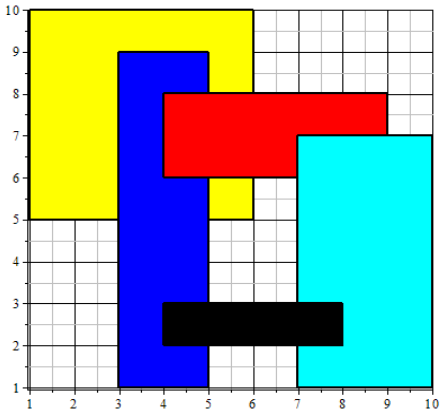

Figure 1 Data Set Bivariate interval censored data

Example.

Suppose given a set rectangles shape of data, such as Figure. 1, with the coordinates $(x_{lj},y_{lj},x_{uj},y_{uj})$ is R_1=(1,10,6,5), R_2=(3,9,5,1), R_3=(4,8,9,6), R_4=(7,7,10,1) and R_5=(4,3,8,2). There are five observations represented by the rectangle R_1, R_2, R_3, R_4 and R_5. As seen in Figure 1 and Figure 2, there are four cliques and that all four are maximal, i.e, M_1= {R_1, R_2, R_3}, M_2= {R_2, R_5}, M_3= {R_3, R_4} and M_4= {R_4, R_5}. The clique graph is displayed in Figure 3.

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). 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 adge (Gentleman and Vandall [4]).

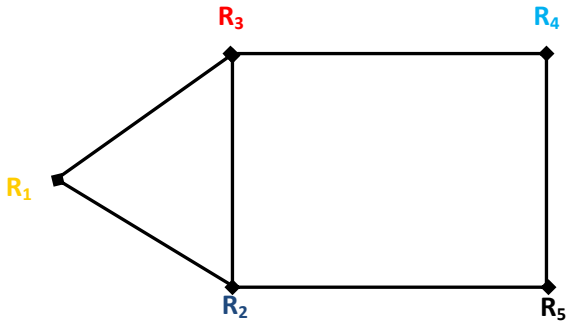

Figure 2. Intersection Graph

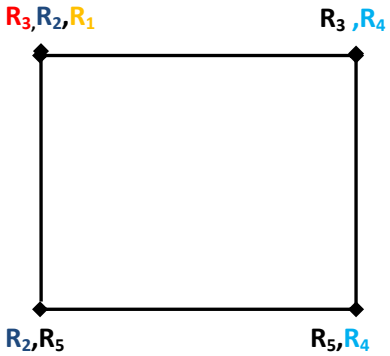

Figure 3. Clique Graph

4. Algorithm of Intersection Graph

Given rectangle with coordinates R_i, i.e. $(x_{1i},x_{2i},y_{1i},y_{2i})$. In 1983 Lee defines the starting vertex (x_{1i}, y_{1i}) at the lower left corner and the second (x_{2i}, y_{2i}) at the upper right corner and assumes that all coordinates of rectangles are different. In this study, the starting vertex (x_{1i}, y_{1i}) is the upper left corner and the (x_{2i}, y_{2i}) is the lower right corner. There is possibility that some individuals have same event times; thus we assume that not all coordinates are different. Algorithm 1 shows some steps for finding intersection graph (see Figure 4 and Figure 5 in Appendix for the flowchart).

4.1 The Steps of Intersection Graph

Given a rectangle R_i, R_j and R_k, for i=1,2,3,...,n, j=1,2,3,...,n, and k=1,2,3,...,n with $i \neq j \neq k$.

1. Input a rectangular coordinates e.g $R_i = (x_{1i},x_{2i},y_{1i},y_{2i})$ starting point (x_{1i},y_{1i}) is the upper left corner and (x_{2i}, y_{2i}) the lower right corner,
2. If intersect R_i with R_j, where $i \neq j$, then the value H_i is 1, otherwise the value H_i is 0. If R_j intersect with R_i then the value of H_j is 0.
3. Repeat Step (2), for R_i and R_k, where $i \neq k$, and for R_j and R_k, where $j \neq k$.

4. If R_i intersect with R_j and R_k, where $i \neq j \neq k$, then the value of H_i is 1, otherwise the value of H_i is 0. If R_j and R_k intersect with R_i then the of H_j and H_k value 0.
5. Repeat until Steps 2 through 4 for all n rectangles.
6. Process will be complete if all possible intersection of R_i, R_j and R_k have been considered.

Tabel 1. Intersection Matrix

	R_1	R_2	R_3	R_4	R_5
H_1	1	1	1	0	0
H_2	0	1	0	0	1
H_3	0	0	1	1	0
H_4	0	0	0	1	1

5. The Steps to Calculate the time Complexity of Algorithms Intersection Graphs :

From Example 1 we obtain an intersection matrix as seen in Table 1. Furthermore we can calculate time complexity of Algorithm 1 as follows:

1. Input a rectangle data, start from 1 to n. It can take at most n steps, so the complexity is $O(n)$.
2. During the initial iteration involving two or more rectangles, we need 1xn steps, so the complexity is $O(n)$.
3. After the iteration process ends, the execution of the occurrence of intersection or the formation of matrix starts from the first row and the first column explaining that the first rectangle intersects with itself. Furthermore the execution on the first row and the second column explains that the first and second rectangles intersect or not. The process continues until the first row and n-th column, that is the first and n-th rectangles intersect or not. This process requires 1xn^2 steps, so the complexity is $O(n^2)$.
4. The process continues to the second row. This process requires 1xn^2 steps, so the complexity is $O(n^2)$.
5. This step continues until all possible intersection of n rectangles are considered. This process requires 1xn^2 steps, so that the complexity is $O(n^2)$. So the time complexity of the algorithm is quadratic.

6. Concluding Remark

This paper has presented algorithm of intersection graph for bivariate censored data by taking the starting vertex on the upper left corner and lower right corner of the rectangle coordinate. The time complexity of this algorithm is $O(n^2)$. This complexity is similar to the result obtained by Lee (1983) in (Maathuis, [7]). This algorithm is a new alternative way to calculate the regions of possible support

for bivariate interval censored data.

REFERENCES

[1] Bogaerts, K. and 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.

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

[3] Gentleman, R. and Vandal, A.C. 2001. Computational algorithms for censored data problems using intersection graphs. *J. Computation and Graphical Statist*,10: 403-421.

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

[5]. Gubler. D.J. 1998. Dengue and Dengue Hemorrhagic Fever. *J.Clinical Microbilogy Reviews*. 11. 480-496

[6]. Liu, X. 2005. *Nonparametric estimation bivariate censored data a discrete* of the *approach.* A thesis submitted to McGill University in partial fulfilment of the requirements degree of Doctor of Philosophy Department of Mathematics and Statistics McGill University, Montreal, Canada

[7] Maathuis, M.H. 2003. *Nonparametric maximum likelihood estimation for bivariate censored data*. A thesis submitted to the Delft University of Technology for the degree of Master of Science/Wiskundig ingenieur.

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

Appendix. Flow Chart of Algorithm 1

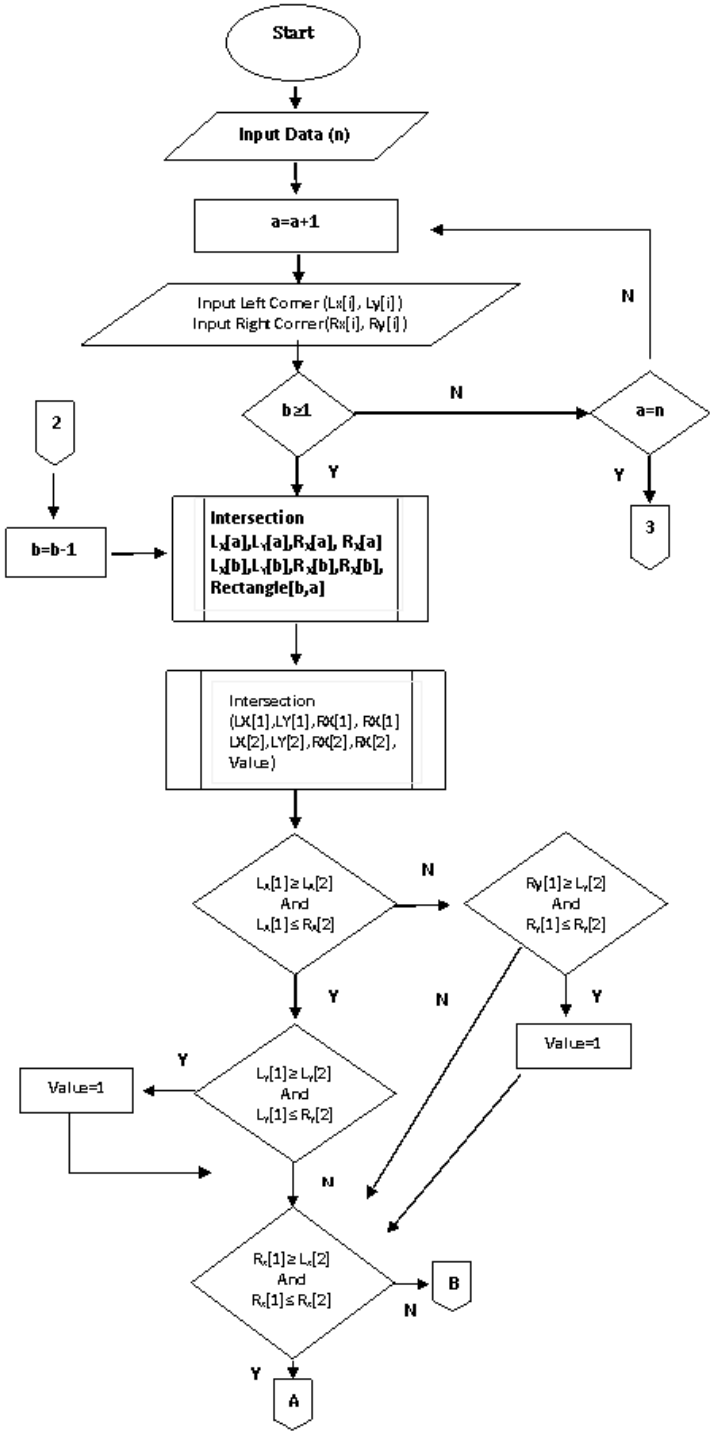

Figure 4

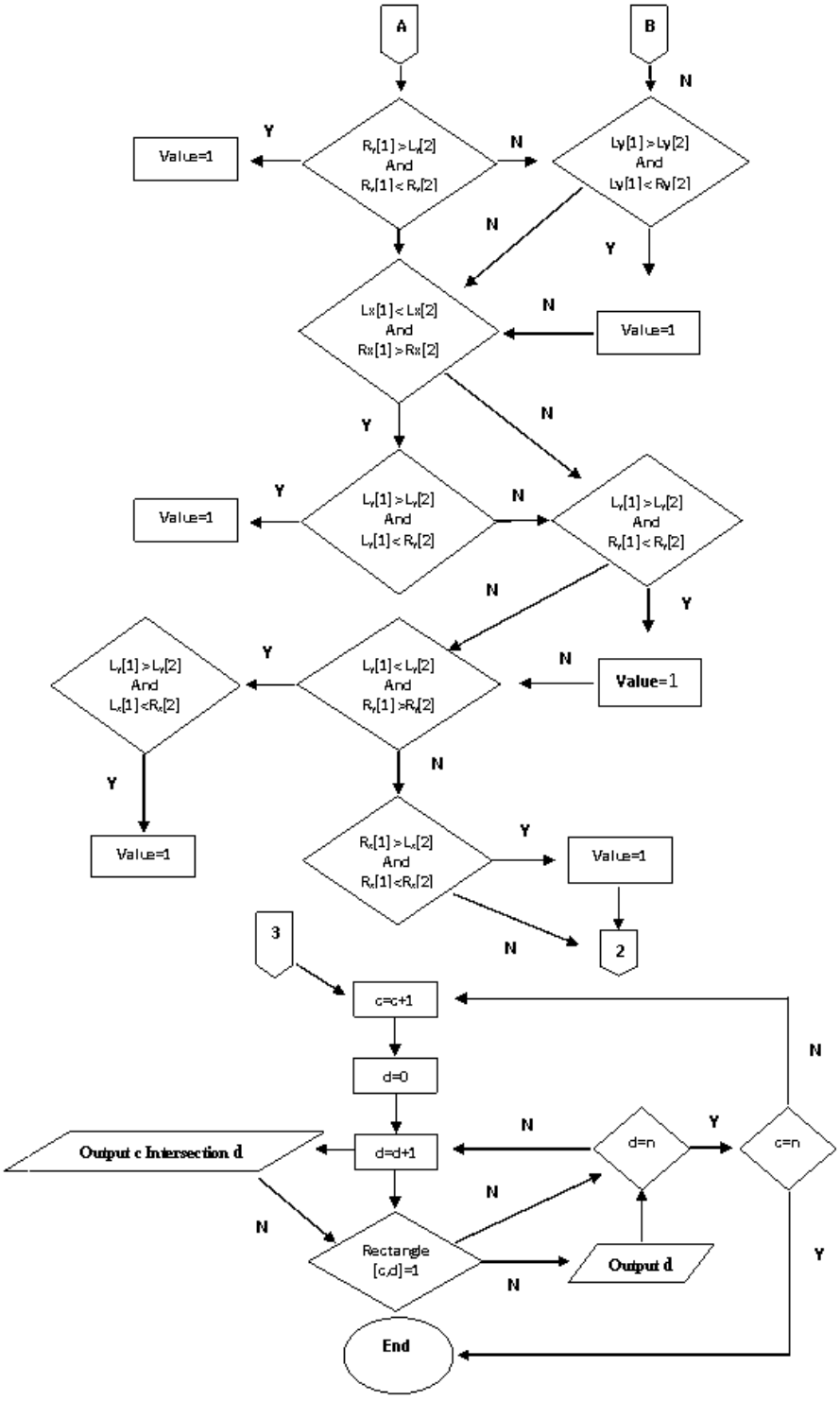
A
B
N
Y
Value=1
$R_y[1] > L_y[2]$ And $R_x[1] < R_x[2]$
$Ly[1] > Ly[2]$ And $Ly[1] < Ry[2]$
N
Y
$Lx[1] < Lx[2]$ And $Rx[1] > Rx[2]$
N
Value=1
Y
N
Y
Value=1
$L_y[1] > L_y[2]$ And $L_y[1] < R_y[2]$
N
$L_y[1] > L_y[2]$ And $R_y[1] < R_y[2]$
N
Y
$L_y[1] > L_y[2]$ And $L_x[1] < R_x[2]$
Y
$L_y[1] < L_y[2]$ And $R_y[1] > R_x[2]$
N
Value=1
N
Y
Value=1
$R_x[1] > L_x[2]$ And $R_x[1] < R_x[2]$
Y
Value=1
N
2
3
c=c+1
d=0
N
d=d+1
N
d=n
Y
c=n
N
Output c Intersection d
N
Rectangle [c,d]=1
N
Output d
Y
End

Figure 5