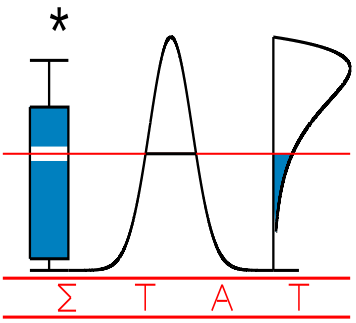
T E C H N I C A L

R E P O R T

0312

A NEW FAST ALGORITHM TO FIND THE REGIONS OF POSSIBLE SUPPORT FOR BIVARIATE INTERVAL CENSORED DATA.

K. BOGAERTS and E. LESAFFRE



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A new fast algorithm to find the regions of possible support for bivariate interval censored data.

Kris Bogaerts[1] and Emmanuel Lesaffre

Katholieke Universiteit Leuven, Biostatistical Centre, Leuven, Belgium

Abstract

The estimation of the non-parametric maximum likelihood estimate (NPMLE) of the bivariate distribution function on interval censored data is a recent topic of research. Among other things, it provides a basic tool for checking a parametric model for the bivariate failure times.
As a first step in the estimation of the NPMLE for bivariate interval censored data, the regions of possible support, i.e. the rectangles with non-zero mass, are calculated. For this step a new fast algorithm is introduced here and compared with two existing algorithms. The advantages of our algorithm will be illustrated on the emergence times of permanent teeth on data from the longitudinal Signal®Tandmobiel study.

Key words: Bivariate survival, interval censored, non-parametric maximum likelihood estimator

1 INTRODUCTION

In dental research, there is interest in examining factors that influence the emergence of permanent teeth. Central to this research is the emergence age of a tooth, i.e. the chronological age of a child at which that tooth appears in the mouth. The pattern, and not merely the timing, of (permanent) tooth emergence is of interest in dental practice for good diagnosis and treatment planning. In forensic dentistry this information is useful for the estimation of the chronological age of a child with unknown birth records.
Among other things, the Signal®-Tandmobiel project provides data on permanent tooth emergence for Flemish children aged 7 to 12 years (see Section 3.3 for further details). As the children were examined annually, many of the emergence times are interval censored. That is, the time T of emergence is only known to lie in an interval $[t_l, t_r]$. For some children, however, the permanent tooth had already emerged at the time of the first visit ($t_l = 0$, left-censored emergence time), while for other children the tooth had not yet emerged at the last visit ($t_r = \infty$, right-censored emergence time). Left- and right-censored data can be seen as special cases of interval censored data. If $t_l = t_r$, the observation is exact.
In the univariate setting, the non-parametric maximum likelihood estimator for interval censored survival data has been studied extensively (e.g. Peto [7], Turnbull [8], Groenenboom and Wellner [6], Böhning et al. [2]). Peto [7] was the first to note that the principle of maximum likelihood estimation leads to a set of intervals $\{(p_j, q_j)\}_{j=1}^{m}$, such that the estimate is constant outside of these intervals and that the mass assigned to an interval is well determined but no information is provided as to how that mass is assigned

[1]Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium, E-mail: Kris.Bogaerts@med.kuleuven.ac.be

within that interval. Peto [7] and Turnbull [8] suggested a simple algorithm to identify these regions of possible mass from the data. Namely, rank the timepoints $\{l_j\}$ and $\{r_j\}$ $(j = 1, \ldots, n)$ in increasing order and keep track of whether the point is a left or a right endpoint. The regions of possible mass are then the intervals with a left endpoint immediately followed by a right endpoint. This observation facilitates the non-parametric estimation of the survivor function considerably.
For bivariate interval censored data, the observed data consist of the rectangles $[l_{1i}, r_{1i}] \times [l_{2i}, r_{2i}]$, $i = 1, \ldots, n$, which are known to contain the unobserved times to the events of interest T_{1i} and T_{2i}, respectively. Note that the rectangles can also be half-planes, half-lines or points. Betensky and Finkelstein [1] (further referred to as B&F) generalized Peto's and Turnbull's argument to bivariate interval censored data. That is, information on the bivariate non-parametric survival function is concentrated in a limited number of rectangles bearing (possibly) non-zero mass. Once the regions of possible support are determined, the non-parametric maximum likelihood estimate (NPMLE) can be estimated by constrained maximization of the likelihood. This can be done using different algorithms like e.g. the EM-algorithm [4] or mixture estimation methods such as the vertex exchange method [2]. Thus for the first step in the calculation of the NPMLE, it is important to determine the regions of possible support efficiently. In the appendix to their paper, B&F provided a simple, but not so efficient, algorithm to calculate the regions of possible support. Based on graph theory, Gentleman and Vandal [5] (further denoted by G&V) proposed a more efficient algorithm and proved that their algorithm has a complexity of at most $O(n^5)$.
Here, a new algorithm is proposed with a complexity of at most $O(n^3)$. For our examples with a sample size of 1000, an acceleration factor of 30 to 45 for the calculation of the regions of possible mass is seen. The three algorithms will be compared based on published data from the literature and for subsamples of various sizes from the Signal®-Tandmobiel study.
In the next section the two existing algorithms and our algorithm are described. Comparisons of our algorithm with the algorithms of B&F and G&V are made in Section 3.

2 ALGORITHMS FOR CALCULATING THE REGIONS OF POSSIBLE SUPPORT

2.1 Review of published algorithms

Here, we review the algorithms of B&F [1] and of G&V [5] for calculating the regions of possible support. The algorithms will be illustrated on the small sample data set introduced in the paper of B&F (see Table 1).
Figure 1 graphically represents the search process of the B&F algorithm. The search process is based on making pairwise intersections of all the observed rectangles and keeping the (non-empty) intersections or the rectangle itself (if there is no intersection with the other rectangles). This procedure is then iterated until no more changes are observed in the list of rectangles of possible support. In Step 0 the observed data are shown. In Step 1, $\binom{6}{2}$ intersections must be made. This results in 6 new rectangles, shown in bold. In Step 2, again $\binom{6}{2}$ intersections must be made. This results in 4 new rectangles. In Step 3 (not shown in the figure), $\binom{4}{2}$ intersections must be made. The same 4 rectangles are found as those identified in Step 2. Since there are no more changes in the set of

Observation	Dimension 1				Dimension 2			
i	l_1		r_1		l_2		r_2	
1	1	(1)	5	(4)	1	(1)	5	(4)
2	4	(3)	8	(7)	4	(3)	8	(7)
3	7	(6)	11	(9)	7	(5)	11	(9)
4	1	(1)	3	(2)	10	(8)	11	(9)
5	6	(5)	10	(8)	3	(2)	7.5	(6)
6	7	(6)	8	(7)	3	(2)	3	(2)

Table 1: Example data set from B&F [1] (ranks of unique values within parentheses, see Section 2.2).

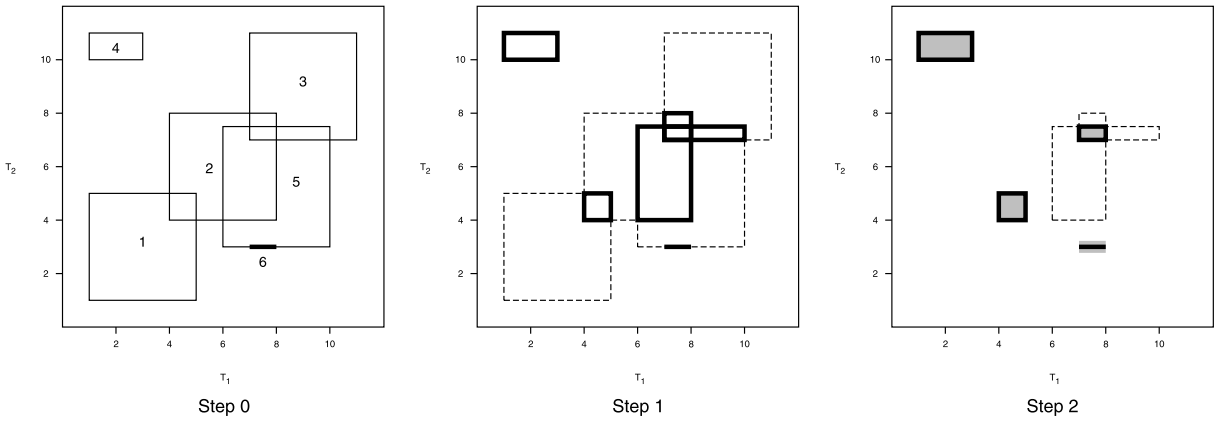


Figure 1: Graphical representation of the algorithm of B&F [1] applied on the example of Table 1.

rectangles of possible support, the algorithm stops. The resulting 4 regions of possible support are shaded. An advantage of this algorithm is the easiness with which it can be implemented. However, considering all pairwise comparisons at each step can yield an overly large number of candidate rectangles when analyzing a large data set.

Using concepts from graph theory, G&V [5] suggested another algorithm. Figure 2 illustrates this algorithm on the same example. The original observations are shown in dashed lines. Expressed somewhat differently, it can be described as follows: in both dimensions, search marginally for the regions of possible support. This can be done using, for instance, Turnbull's algorithm on all left and right endpoints. The regions are represented on both axes of Figure 2 by the space between the arrows. For each interval of possible support, determine the bivariate observations (i.e. rectangles) that contain that specific interval. The collection of these observations for that interval is called a maximal clique. For e.g. the first dimension, the maximal cliques are {1,4}, {1,2} and {2,3,5,6}. The final regions of possible support are a subset of the intersections of all these maximal cliques from both dimensions. The algorithm searches this set of intersections in a straightforward manner. The final regions of possible support are shaded on the figure. G&V proved that their algorithm has a complexity of at most $O(n^5)$. More details on the algorithm and the terminology from graph theory can be found in the original publication [5].

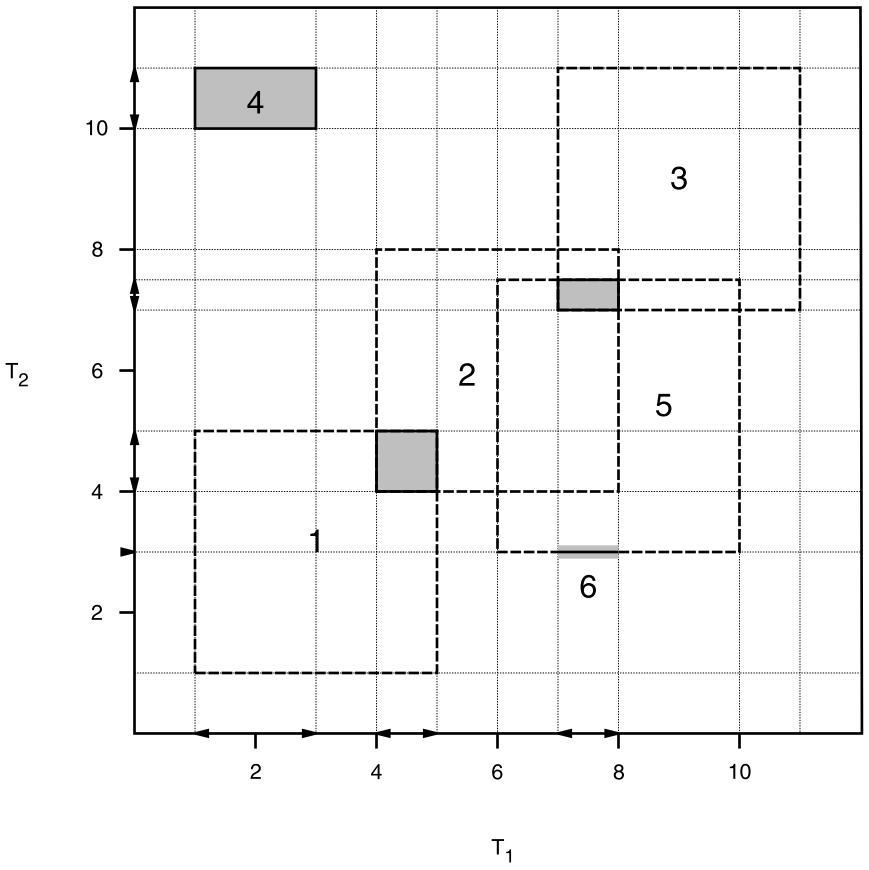


Figure 2: Graphical representation of the algorithm of G&V [5] applied on the example of Table 1.

2.2 Description of a new algorithm

Figure 3 is a graphical representation of our algorithm, which can be described as follows: for each distinct endpoint (left or right) on the Y-axis (say $Y = j$), determine the regions of possible support (e.g. with Turnbull's algorithm) on the intervals that are obtained by intersecting the observed rectangles (which are displayed in dashed lines) with the horizontal line $Y = j$. This results in a set of endpoints (l, r) along the X-axis with possible non-zero mass for each left or right endpoint on the Y-axis. For e.g. $Y = 3$, we find 2 regions of possible support, namely [1,5] and [7,8]. The resulting regions of possible mass (stratified for each value of the other dimension) are displayed in Figure 3 by the space between the arrows. Do this similarly for the left and right endpoints on the X-axis. So for e.g. $X = 4$, we find one possible region of support, namely [4,5]. Finding the regions of possible mass for the bivariate case is then done by trying to construct rectangles out of the univariate regions of support we just calculated. Taking for instance interval [7,8] for $Y = 7$ as a possible basis for a rectangle of possible support, we have to check that from both endpoints (left and right) of this interval a vertical (univariate) region of possible support of equal length can be found. This is the case, namely [7,7.5] for $X = 7$ and $X = 8$. Finally, we have to check whether there is a horizontal (univariate) region of support that connects the upper endpoints of the vertical regions of possible support (i.e. closes the rectangle we are constructing). Here, this is done by [7,8] for $Y = 7.5$. If within the rectangle we constructed ($[7, 8] \times [7, 7.5)]$) no other such rectangle can be found, then we have a region of possible support. The latter requirement is necessary for excluding false rectangles of possible support. The necessity of this requirement is easily understood by considering an artificial example where one observation lies completely inside another one. On the other hand, if we had taken for instance interval [1,5] for $Y = 3$ as a basis for the rectangle of possible support, we notice that on the right endpoint no vertical region of possible support is present. Therefore interval [1,5] for $Y = 3$ cannot be the basis of a region of possible support. More formally expressed, the regions of possible mass for the bivariate case are determined by rectangles that are build out of regions of possible mass from the univariate stratified process and have no other such rectangle completely inside it. For the example data, four such rectangles can be found. They are shaded on the figure. (Note that for the region of possible support $[7, 8] \times [3, 3]$ the vertical arrows cannot be displayed). At first sight, an obvious modification of our algorithm was suggested by one of the referees. More specifically, one could condition only on the endpoints of the marginal regions of support in a specific dimension. Unfortunately, this approach would yield incorrect results as is illustrated in the following example. From the two observations [1,2]x[1,3] and [3,4]x[2,4], it is easily seen that the 2 observations are disjoint and therefore both should be a region of possible support. However, if one conditioned in the second dimension on the endpoints of the marginal region of support [2,3], it would not be possible to reconstruct the original observations as regions of possible support.

A SAS®(version 8.2) macro for the new algorithm can be obtained from
http://www.kuleuven.ac.be/biostat/research/software.htm.

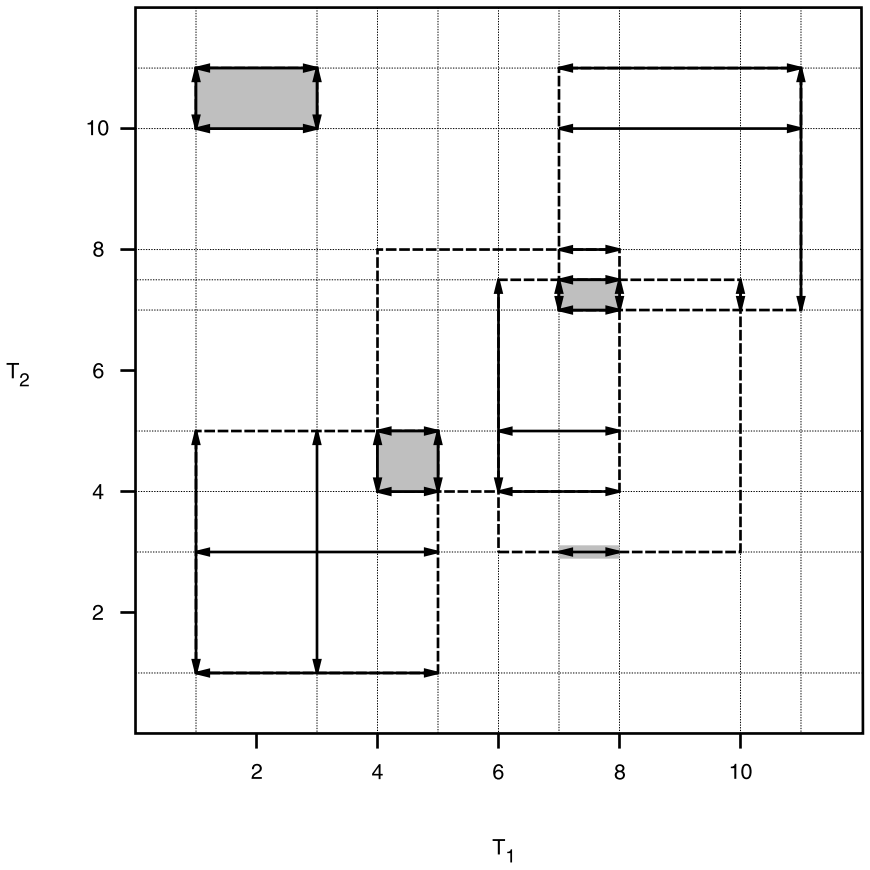


Figure 3: Graphical representation of the new algorithm applied on the example of Table 1.

3 COMPARISON WITH EXISTING ALGORITHMS

Our algorithm is compared with the algorithms of B&F and G&V both theoretically and by applying them to various data sets. For the purpose of comparison, the three algorithms were implemented in SAS®IML version 8.2. The comparison is done on a PC running Windows XP Home Edition with a 2 GhZ processor and 512 MB RAM. For each data set, the time to compute the matrix (also called clique matrix by G&V) that is needed in the second step of calculating the NPMLE is reported.

3.1 Theoretical comparison

Betensky and Finkelstein did not provide a complexity analysis of their algorithm. In practice, such an analysis seems difficult to perform. The algorithm of B&F is not suitable to be used with data sets containing a large number of distinct observation endpoints. By making all pairwise comparisons, the number of candidates for a region of possible support is inflated tremendously during the search process. For instance, for the first data set of size 50 from the Signal®-Tandmobiel study (see Section 3.3), the algorithm creates in the first step $\binom{50}{2} = 1\,225$ intersections leading to 566 new rectangles. Therefore, in the second step $\binom{566}{2} = 159\,895$ intersections are created. In the third step, $\binom{696}{2} = 241\,860$ intersections are created. Eventually the number of intersections in each step will decrease again. As can be seen from this example, the number of intersections to be made in each step can grow very rapidly. As a result, the algorithm will require a large computation time.

By using graph theory, G&V have developed a nice framework to study the problem of finding the NPMLE of the distribution function for survival data. They proved that their algorithm has a worst-case complexity of $O(n^5)$.

Instead of looking for the maximal cliques from which the regions of possible mass can be determined, our approach directly searches for the regions of possible mass. While doing so we were able to reduce the worst case complexity from $O(n^5)$ to $O(n^3)$. This can be understood as follows: there are at most $2n$ different endpoints (left and right) in each dimension. Sorting all endpoints can be done in $O(n \log n)$ by a variety of algorithms. Performing Turnbull's algorithm for each stratified value (at most $2 \times 2n$ levels) can also be done in $O(n)$. Over all stratified values this therefore yields a complexity of $O(n^2)$. For determining the regions of possible support, all possible rectangles must be checked. This can be done at a worst-case complexity of $O(n^3)$ (see the Appendix for more details).

The time needed for the creation of the clique matrix from the regions of possible support does not undo the gain that is obtained by searching for the regions of possible support directly. In addition, this gain is independent of the software implementation of the algorithms as our algorithm programmed in S-plus showed about the same gain as the original R program of Gentleman and Vandal (kindly provided to us).

A disadvantage of the new algorithm compared to the algorithm of G&V is that it requires more memory. In the worst case, several matrices of size $2n$ are needed. As n becomes large, this could become too demanding.

The algorithm of G&V is quite easily extended to higher dimensions. Our algorithm can also be extended to higher dimensions but the extension is less transparent, certainly for more than 3 dimensions. Further research is also needed to determine which of the two algorithms would perform best in higher dimensions.

3.2 Comparison based on published data

The first comparison is based on the data set presented in B&F [1]. It concerns data from a natural substudy of a comparative trial from the AIDS Clinical Trials Group (protocol ACTG 181). Two hundred and four patients were scheduled for clinic visits during follow-up and data were collected until either of the following two events happened: shedding of cytomegalovirus (CMV) in the urine and blood, or colonization of Mycobacterium Aviam Complex (MAC) in the sputum or stool.
All three algorithms found the same 32 regions of possible support. The original algorithm of B&F took 32 seconds, the algorithm of G&V took 0.07 seconds and our algorithm 0.17 seconds.

3.3 Comparison based on tooth emergence data

A second comparison is done on subgroups of different sizes (50, 100, 500 and 1000) from the Signal®-Tandmobiel study. The Signal®-Tandmobiel study is a prospective longitudinal survey, which collected dental, oral hygiene and dietary data from a representative sample of Flemish children born in 1989. They represent about 7% of the total population of Flemish children of the same age. In total 4468 children, 2153 (48.2%) girls and 2315 (51.8%) boys, were followed. The children were examined annually on pre-scheduled visits (from the age of 7 to the age of 12) by trained dentist-examiners in a mobile dental clinic on the school premises. The time between visits ranged from 0.4 to 1.6 year, and had a median at 0.9 year. A more elaborate description of the Signal®-Tandmobiel project can be found in Vanobbergen et al. [9].
Tooth emergence was recorded at each examination by direct inspection. Every individual permanent tooth was scored according to its clinical eruption stage (adapted from Carvalho et al. [3]). For this analysis however, the status of tooth eruption was dichotomized: not emerged versus emerged.
We have taken three pairs of teeth which differ in the amount of perceived correlation (not determined) between two interval censored emergence ages. Namely for data set 1 the maxillar central incisors, for data set 2 the maxillar first premolars and for data set 3 the maxillar right first premolar and the mandibular left canine were taken. The emergence ages for data sets 1 and 2 are expected to be highly correlated. For data set 3 the perceived correlation is much lower.
The algorithm of B&F was tried out on only one of the smallest data sets of size 50. After more than 21 hours, the computations were stopped.
The results for our algorithm and that of G&V are summarized in Table 2. The two algorithms always found the same regions of possible support. For the smaller data sets of sizes 50 and 100, the algorithm of G&V was slightly faster. However, for the larger data sets, our algorithm outperformed the algorithm of G&V considerably. For the largest sample size tested, i.e. 1000, an acceleration factor of 30 to 45 could be seen. This can be explained by the fact that our algorithm conducts a search for the regions of possible support after applying the stratified search for the one-dimensional regions of possible support. When many regions of possible support are present, the combinatorial algorithm

Sample size	Data set	number of regions with possible support	Gentleman & Vandal[5]	New algorithm
50	1	17	0.06s	0.17s
	2	32	0.09s	0.18s
	3	38	0.10s	0.20s
100	1	73	0.57s	1.12s
	2	85	0.85s	1.12s
	3	112	0.99s	1.29s
250	1	359	38s	15s
	2	368	43s	14s
	3	499	60s	16s
500	1	777	10m21s	1m23s
	2	1096	16m26s	1m27s
	3	2133	31m09s	2m19s
1000	1	2742	4h16m41s	7m30s
	2	4246	8h28m28s	11m21s
	3	7838	15h06m01s	20m22s

h=hours, m=minutes, s=seconds.

Table 2: Comparison of computation times of the new algorithm with that of G&V for several subsamples from the Signal®-Tandmobiel project.

of G&V slows down considerably because for a new candidate the algorithm loops always through the list of already found solutions in order to adapt the list. It is the most time consuming process of their algorithm.

4 FINAL REMARKS

Throughout the manuscript, the convention of closed intervals has been used. However one must pay attention to common inspection times as they may cause computational problems. For instance, the intervals [2,3] and [3,4] would yield [3,3] as possible region of support, whereas (2,3] and (3,4] would yield both intervals as regions of possible support. To solve this problem, one could add (or subtract) a small enough value (e.g. 0.0001) to the left (or right) endpoint. One must make sure that no other inspection time lies between the original inspection time and the artificially created one.
The calculation of the NPMLE in the case of bivariate interval censored data consists of 2 parts. The first part involves the determination of the regions of possible support and the second part the maximization of the likelihood. If one looks for an efficient algorithm to calculate the NPMLE, then both parts should be calculated as efficiently as possible. The maximization of the likelihood has already been studied extensively (e.g. Peto [7], Turnbull [8], Böhning et al. [2]), but mostly in the univariate case. Some algorithms like for instance those from mixture methods (Böhning et al. [2]) can be used without modification in the bivariate case. Also a quasi-Newton maximization algorithm (for example PROC NLP from SAS®version 8.2) could be used. For the calculation of the regions of possible support for the bivariate case, not much research has been performed yet. The two algorithms reviewed in this paper were the only published algorithms that we have

found. One of the referees pointed out another algorithm to find the maximal cliques by Tsukiyama et al. [10]. However, it can been shown that the worst case complexity of this algorithm is also $O(n^5)$.
By using graph theory, G&V have developed a nice framework to study the problem of finding the NPMLE of the distribution function for survival data. They proved that when the row-rank of this matrix is equal to the number of regions of possible support, then the NPMLE is unique. However they note that row-rank deficiency is necessary but not sufficient for non-uniqueness. Clearly this is an easy tool to determine the uniqueness of the NPMLE. The algorithm of G&V provides the cliques matrix directly. For our algorithm, one still has to calculate the cliques matrix from the regions of possible support. However, the extra time required to do so is negligible compared to the gain that is made by directly searching for the regions of possible support. Therefore, the tool for determining the uniqueness of the NPMLE can also be applied after our algorithm.
In conclusion, our algorithm provides a new and fast way to calculate the regions of possible support for large data sets of bivariate interval censored data.

Acknowledgements
The second author was partly financially supported by Research Grant OT/00/35, Catholic University Leuven.
Both authors acknowledge support from the Interuniversity Attraction Poles Program P5/24 - Belgian State - Federal Office for Scientific, Technical and Cultural Affairs.
The Signal-Tandmobiel®project comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Working Group Oral Health Promotion and Prevention, Flemish Dental Association; Dental School, University Ghent), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, University Leuven), K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References

[1] R.A. Betensky and D.M. Finkelstein. A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18:3089–3100, 1999.

[2] D. Böhning, P. Schlattmann, and E. Dietz. Interval censored data : A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika*, 83:462–466, 1996.

[3] J.C. Carvalho, K.R. Ekstrand, and A. Thylstrup. Dental plaque and caries on occlusal surfaces of first permanent molars in relation to stage of eruption. *J Dent Res*, 68(5):773–779, 1989.

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[5] R. Gentleman and A.C. Vandal. Computational algorithms for censored-data problems using intersection graphs. *Journal of computational and graphical statistics*, 10(3):403–421, 2001.

[6] P. Groeneboom and J.A. Wellner. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, 1992.

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

[8] B.W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of The Royal Statistical Society*, 38:290–295, 1976.

[9] J. Vanobbergen, L. Martens, E. Lesaffre, and D. Declerck. The Signal-tandmobiel® project, a longitudinal intervention heath promotion study in Flanders (Belgium) : baseline and first year results. *Eur J Paediatr Dent*, 2:87–96, 2000.

[10] S. Tsukiyama, M. Ide, H. Ariyoshi and I. Shirakawa. A new algorithm for generating all the maximal independent sets. *SIAM J. Comput.*, 6(3):505-517, 1977.

Appendix

Complexity of the new algorithm.
As described in Section 3.1, all operations up to the conditional scanning of the observations can be done with a complexity of at most $O(n^2)$. The determination of the bivariate regions of possible support from the conditional regions of possible support (these are intervals) can be done in $O(n^3)$. This can be seen as follows: to find the bivariate regions of possible support, we have to look for rectangles of which the sides are intervals that we found by conditionally performing Turnbull's algorithm. First we look for an interval of possible support in the first dimension. Once such an interval is found, we look whether its begin- and endpoint are also beginpoints of an interval of possible support in the second dimension. If so, we have to determine the endpoints of these 2 intervals in the second dimension. For each interval, we have to perform on average $\frac{m_2 \cdot (m_2 - 1)}{2 \cdot m_2}$ steps to do so. In addition, the endpoints of both these intervals must have the same value in the second dimension. Further, the endpoint of the interval on the left-hand side should be the start of an interval of possible support in the first dimension and the endpoint from this interval must coincide with the endpoint of the interval on the right-hand side. For this check, on average $\frac{m_1 \cdot (m_1 - 1)}{2 \cdot m_1}$ steps are required. When a rectangle that fulfills the above requirements is found, we have to check in addition that within this rectangle no other (internal) rectangle with the same properties is located. We can do this by checking for each value between the left-hand and right-hand side of the rectangle whether within the rectangle another begin- or endpoint can be found. This can be done for each value in on average $\frac{m_2 \cdot (m_2 - 1)}{2 \cdot m_2}$ steps. If we loop through the matrix in $m_1 \cdot m_2$ steps, we can subdivide these steps in a steps where we should check for a rectangle of possible support and in b steps where we have to check for an internal rectangle, with $a + b \leq m_1 \cdot m_2$. Therefore we need only $a \cdot (2 \cdot \frac{(m_2 - 1)}{2} + \frac{(m_1 - 1)}{2}) + b \cdot \frac{(m_2 - 1)}{2}$ steps to find all bivariate regions of possible support. Because m_1, $m_2 \leq 2n$, we can look for the bivariate regions of possible support in a complexity of at most $O(n^3)$.