Interval Censored Data: A Note on the Nonparametric Maximum Likelihood Estimator of the Distribution Function

Dankmar Bohning, Peter Schlattmann, Ekkehart Dietz

Biometrika, Volume 83, Issue 2 (Jun., 1996), 462-466.

Biometrika (1996), **83**, 2, *pp.* 462–466
Printed in Great Britain

Interval censored data: A note on the nonparametric maximum likelihood estimator of the distribution function

BY DANKMAR BÖHNING, PETER SCHLATTMANN AND EKKEHART DIETZ

Department of Epidemiology, Free University Berlin, Augustastraße 37, 12203 Berlin, Germany

SUMMARY

Gentleman & Geyer (1994) discuss the analysis of interval censored data and present results based on standard convex optimisation theory. Here, this problem is viewed from the perspective of a mixing problem of indicator functions. Using an analogy with the problem of mixture distributions a variety of results are easily derived, including a characterisation theorem for the maximum likelihood estimator and various reliably convergent algorithms. Software for the analysis of mixture distributions can be used to find the nonparametric estimate of the distribution function of the interval-censored survival time. Examples are provided to demonstrate the theory.

Some key words: C.A.MAN; Interval censoring; Mixture model; Nonparametric estimator of survival time; Vertex exchange algorithm.

1. INTRODUCTION

The problem and notation are as in Gentleman & Geyer (1994). The unobserved survival time X is supposed to arise from a distribution F_0 and to lie in an observed open interval (L_i, R_i), corresponding to the last inspection time prior to the life event and the first inspection time after the life event. Thus, the observable data are open intervals $I_1, \dots, I_n$, with $I_i = (L_i, R_i)$ for $i = 1, \dots, n$. The likelihood is defined as $\prod_i \{F_0(R_i-) - F_0(L_i)\}$, where the product is taken from 1 to n. Let $\{s_j\}_{j=1}^m$ denote the unique ordered elements of $\{0, \{L_i\}_{i=1}^n, \{R_i\}_{i=1}^n\}$. Note that $R_i = \infty$ is possible. If α_{ij} denotes the indicator of the event $(s_{j-1}, s_j) \subseteq I_i$, $p_j = F_0(s_j-) - F_0(s_{j-1})$, then, as outlined in Gentleman & Geyer (1994), the likelihood can be written as

$$L = \eta_1 \times \eta_2 \times \dots \times \eta_n, \tag{1}$$

$\eta_i := (\alpha_{i1} p_1 + \dots + \alpha_{im} p_m)$ being the ith contribution to the likelihood. A general introduction to nonparametric maximum likelihood estimation for censored data can be found in Groeneboom & Wellner (1992). Finding the maximum likelihood of F_0 becomes the problem of maximising $l := \log L$ over $p_1, \dots, p_m$, subject to $p_j \geqslant 0$ for $j = 1, \dots, m$ and $p_1 + \dots + p_m = 1$. The problem is to maximise $l(p)$ in the finite dimensional probability simplex

$$\Delta := \{p := p_1 e_1 + \dots + p_m e_m \mid p_j \geqslant 0;\ j = 1, \dots, m,\ p_1 + \dots + p_m = 1\},$$

with e_j being the vector having only 0's except for one 1 at the jth position. Mainly for numerical purposes, we frequently consider the log-likelihood in its geometric mean version $\bar{l} := n^{-1} l(p)$.

2. RESULTS

2·1. *General*

The purpose of this note is to point out and exploit some analogies between the problem at hand and the problem of finding the nonparametric maximum likelihood estimator of a mixing distribution. In a variety of papers, Böhning (1982), Lindsay (1983), Böhning (1989), Böhning, Schlattmann & Lindsay (1992) and Böhning (1995) developed a theory for the nonparametric maximum likelihood estimator of the mixing distribution P giving weights $p_1, \ldots, p_m$ to parameter values $\theta_1, \ldots, \theta_m$ for some parametric density $f(x, \theta)$. The problem of finding the maximum likelihood estimator of the mixing distribution is to maximise $L^* := \prod_i \{\sum_j f(x_i, \theta_j) p_j\}$ over $p_1, \ldots, p_m$ if $\theta_1, \ldots, \theta_m$ are considered known. This clearly points the analogy to (1). We restate the relevant theorems.

2·2. *Characterising the maximum likelihood estimate*

As pointed out by Gentleman & Geyer (1994), $l(p)$ is a concave function on Δ. Define the directional derivative

$$\Phi(p, q) = \lim_{\alpha \to 0} \{l((1-\alpha)p + \alpha q) - l(p)\}/\alpha$$

for any p in Δ and any direction q. Note that $\Phi(p, q) = \nabla l(p)^{\mathrm{T}} q - n$ is a linear function of the direction q. Here $\nabla l(p)^{\mathrm{T}} = (d_1, \ldots, d_m)^{\mathrm{T}}$ with $d_k = \partial l/\partial p_k = \sum_i (\alpha_{ik}/\eta_i)$. As Gentleman & Geyer (1994) point out, d_k is the sum of $1/\eta_i$ for all individuals whose intervals, I_i, intersect the interval (s_{k-1}, s_k). Because l is concave $\Phi(p, q) \geqslant l(q) - l(p)$, and hence

$$\sup_{q \in \Delta} \Phi(p,q) \geqslant l(\hat{p}) - l(p). \tag{2}$$

Since $\Phi(p, q) = \sum_j q_j \Phi(p, e_j)$, where e_j is the jth vertex of Δ, the inequality (2) becomes

$$D(p) := \sup_{1 \leqslant k \leqslant m} \sum_i \frac{\alpha_{ik}}{\eta_i} - n \geqslant l(\hat{p}) - l(p). \tag{3}$$

Clearly, if $D(\hat{p}) = 0$, $\hat{p}$ maximises l globally. This is known as the general mixture maximum likelihood theorem (Lindsay, 1983), stating that $D(\hat{p}) = 0$ if and only if $\hat{p}$ maximises l globally in Δ. Moreover, if $\hat{p}_k > 0$, then $\Phi(\hat{p}, e_k) = 0$, in other words $\sum_i (\alpha_{ik}/\eta_i) = n$. For the geometric mean version of the likelihood the theorem takes the form: $\bar{d}_k = \partial \bar{l}/\partial p_k \leqslant 1$ for all $k = 1, \ldots, m$, and $\bar{d}_k = 1$ for those k with p_k strictly positive. We demonstrate this theorem with the example given in Gentleman & Geyer (1994, § 4).

Example 1. Let the data consist of the six intervals (0, 1], (1, 3], (1, 3], (0, 2], (0, 2], (2, 3]. This leads to the following 6 × 3 matrix of indicators:

$$A = (\alpha_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and the likelihood is

$$L = \prod_{i=1}^{6} (\alpha_{i1} p_1 + \alpha_{i2} p_2 + \alpha_{i3} p_3).$$

We have that

$$d_1 = 1/p_1 + 2/(p_1 + p_2), \quad d_2 = 2/(p_2 + p_3) + 2/(p_1 + p_2), \quad d_3 = 2/(p_2 + p_3) + 1/p_3.$$

From this it is clear that $\hat{p} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^{\mathrm{T}}$ is the maximum likelihood estimate, since $d = (6, 6, 6)^{\mathrm{T}}$ meets the condition of the mixture maximum likelihood theorem. The point $(\frac{1}{2}, 0, \frac{1}{2})^{\mathrm{T}}$ is clearly not the maximum likelihood estimate, since the partial derivative $d_2 = 8 > 6$ at this point. Thus, the likeli-

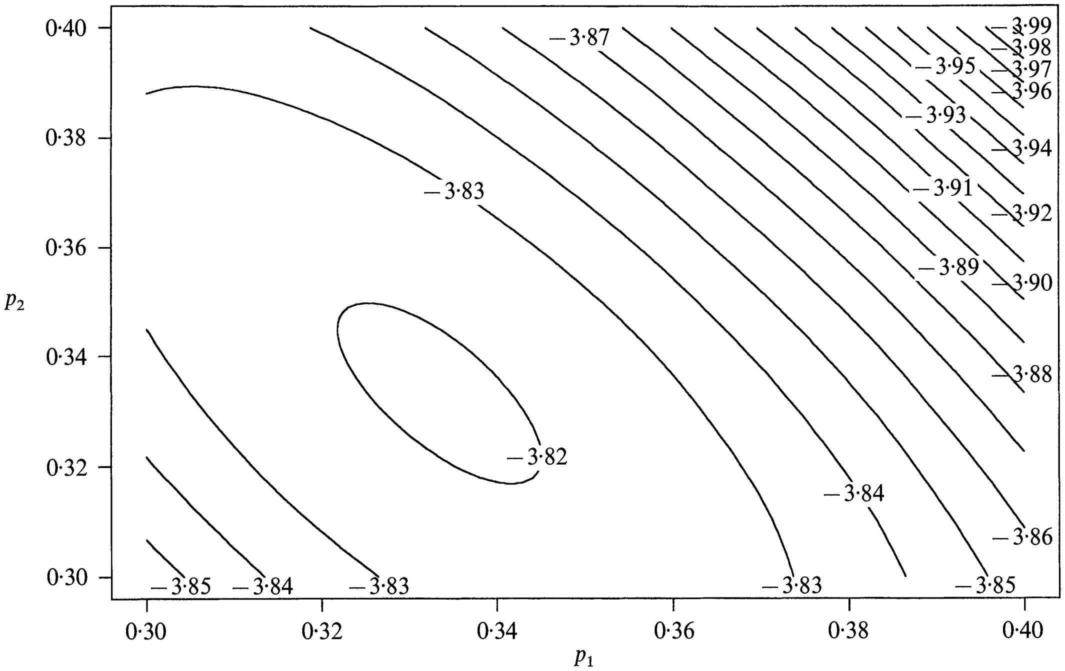

Fig. 1. Contour plot of the log-likelihood, $\log(p_1)+2\log(1-p_1)+2\log(p_1+p_2)+\log(1-p_1-p_2)$, for Example 1.

hood can be increased by moving in the direction of the second vertex e_2. This can be seen also by looking at the contour plot of the log-likelihood when it is written as a function of p_1 and p_2 alone:

$$\prod_{i=1}^{6}\{\alpha_{i1}p_1+\alpha_{i2}p_2+\alpha_{i3}(1-p_1-p_2)\}=p_1(1-p_1)^2(p_1+p_2)^2(1-p_1-p_2).$$

The log-likelihood in this case is

$$l(p)=\log(p_1)+2\log(1-p_1)+2\log(p_1+p_2)+\log(1-p_1-p_2).$$

The contour plot in Fig. 1 clearly encircles the maximum likelihood estimate $(\frac{1}{3},\frac{1}{3},\frac{1}{3})^{\mathrm{T}}$.

Example 2. The second example is more challenging. The data are from Finkelstein & Wolfe (1985) and consist of time intervals in which cosmetic deterioration for early breast cancer patients treated with radiotherapy occurred in 46 individuals. The intervals are listed in Table 1 in Gentleman & Geyer (1994). To save space, the corresponding indicator (46×14)-matrix $A=(\alpha_{ij})$ is not given here, but is obtainable from the authors if desired. For these data, it is not possible to find the maximum likelihood estimate in closed form. Algorithmic approaches have to be applied.

2·3. *Algorithms*

A simple and reliable algorithm is a version of the EM algorithm (Dempster, Laird & Rubin, 1977) which in this case reduces to the iteration p^{EM} with jth component $p_j^{\mathrm{EM}}=\bar{d}_j p_j$. The EM algorithm has the monotonicity property $l(p^{\mathrm{EM}})\geqslant l(p)$, if p is the current iterate, which ensures reliable convergence to the global maximum for arbitrary starting values with components all strictly positive. However, for flat likelihood surfaces, as in this case, its convergence can be rather slow. In addition, many of the p_j might be zero, and, as pointed out by Gentleman & Geyer (1994), the EM algorithm takes extra time to identify these; in Example 2, 6 of 14 weights are zero. Gentleman & Geyer (1994) suggest restarting the EM algorithm with the candidate weights set to zero. This will speed up the convergence, but it will also run the risk of falsely setting a weight to zero, such as p_2 in Table 2 of Gentleman & Geyer (1994). Since the EM algorithm does not change zero weights, it will only maximise the log-likelihood in the corresponding sub-simplex. Indeed, if we consider p given in column 7 of Table 2 in Gentleman & Geyer (1994), that is $p_2 = 0$ in particular, we find that $\bar{d}_2 = 1\cdot1327$, indicating that the likelihood can be increased by putting more weight on the second vertex.

In mixture models a variety of algorithms have been developed (Böhning, 1989, 1995), the useful one being the vertex-exchange algorithm. It is defined by the iteration $p + \alpha p_{\min}(e_{\max} - e_{\min})$, where 'max' and 'min' are integer indices between 1 and m, $p_{\min}$ corresponds to the weight of the vertex $e_{\min}$, and α is a monotonic or optimal step-length in the closed interval $[0, 1]$; see Böhning (1989, 1995) for details. The name vertex-exchange algorithm stems from its property of finding a zero weight in one step ($\alpha = 1$). Since in this context the possibility of many weights being zero arises, the vertex-exchange method should be quite useful here as well. The index 'max' can be determined as $d_{\max} = \max\{d_j \mid j = 1, \ldots, m\}$, whereas 'min' can be found as that index for which the partial derivative is minimised under those with positive weight p_j: $d_{\min} = \min\{d_j \mid j = 1, \ldots, m \text{ and } p_j > 0\}$.

The statistical package C.A.MAN has been developed for the analysis of mixture models (Böhning et al., 1992) utilising various algorithms including EM, the vertex-exchange algorithm, and others. This package computes the nonparametric estimator of the mixing distribution for various densities, including the option of the known density case (Titterington, Smith & Makov, 1985, p. 152). If one uses this option, e.g. if one thinks of the matrix $A = (\alpha_{ij})$ as $F = (f(x_i, \theta_j))$, with $\theta_1, \ldots, \theta_m$ known, then C.A.MAN can be used directly without further modification, and the whole variety of powerful algorithms is available for this setting of interval censored data.

Table 1. *Weights and gradient at maximum likelihood estimate* ($\max_{1 \leqslant k \leqslant m} \bar{d}_k \leqslant 0\cdot000001$), $l(\hat{p}) = -58\cdot06002$

$\bar{d}_k$	p_k	k	$\bar{d}_k$	p_k	k
1·0000	0·0463	1	0·7965	0·0000	8
1·0000	0·0334	2	1·0000	0·0818	9
1·0000	0·0887	3	0·7713	0·0000	10
1·0000	0·0708	4	0·9377	0·0000	11
0·4722	0·0000	5	1·0000	0·1209	12
0·8337	0·0000	6	0·9394	0·0000	13
1·0000	0·0926	7	1·0000	0·4656	14

Example 2 (*cont*). The maximum likelihood estimate given in Table 1 could be identified using any of the algorithms available in C.A.MAN. Note that, for any algorithm, the iteration was stopped if $\max_{1 \leqslant k \leqslant m} \bar{d}_k \leqslant 0\cdot000001$, thus guaranteeing that $l(\hat{p}) \leqslant l(p^{\text{iterate}}) + 10^{-6} \times n$ by inequality (3). The EM algorithm needs a couple of hundred steps to reach this bound, whereas the vertex-exchange method identifies the nonzero weights in only a few steps. The estimate given in Table 1 corresponds to that given by Gentleman & Geyer (1994), as it should. However, it is interesting to observe the deviation in the fourth digit for weight 12 and weight 14, which is given as $p_{12} = 0\cdot1206$ and $p_{14} = 0\cdot4658$ by Gentleman & Geyer (1994). This is probably due to the fact that the EM algorithm was stopped too early: we reached a similar value when using the stopping rule $\max_{1 \leqslant k \leqslant m} \bar{d}_k \leqslant 0\cdot0001$. This phenomenon is observed quite frequently in mixture models (Titterington et al., 1985, p. 90).

ACKNOWLEDGEMENT

This research has been done under the support of the German Research Foundation. The authors are grateful to the Editor, the Associate Editor, and two referees for their constructive comments leading to an improved version of the paper.

REFERENCES

BÖHNING, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.* **10**, 1006–8.

BÖHNING, D. (1989). Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms. *Biometrika* **76**, 375–83.

BÖHNING, D. (1995). A review of reliable maximum likelihood algorithms for the semiparametric mixture maximum likelihood estimator. *J. Statist. Plan. Inf.* **47**, 5–28.

BÖHNING, D., SCHLATTMANN, P. & LINDSAY, B. G. (1992). Computer assisted analysis of mixtures (C.A.MAN): Statistical algorithms. *Biometrics* **48**, 283–303.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

FINKELSTEIN, D. M. & WOLFE, R. A. (1985). A semiparametric model for regression analysis of interval censored failure time data. *Biometrics* **41**, 933–45.

GENTLEMAN & GEYER, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* **81**, 618–23.

GROENEBOOM, P. & WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhäuser.

LINDSAY, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11**, 86–94.

TITTERINGTON, D. M., SMITH, A. F. M. & MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

[*Received April* 1995. *Revised September* 1995]