Maximum Likelihood for Interval Censored Data: Consistency and Computation

By

Robert Gentleman[1] and Charlés J. Geyer[2].

Technical Report No. 588
School of Statistics
University of Minnesota
February 25, 1993

[1] supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada

[2] Research supported in part by grant DMS-9007833 from the National Science Foundation

Abstract

Standard convex optimization techniques are applied to the analysis of interval censored data. These methods provide easily verifiable conditions for the self-consistent estimator proposed by Turnbull (1976) to be a maximum likelihood estimate and for checking whether the maximum likelihood estimate is unique. A sufficient condition is given for the almost sure convergence of the maximum likelihood estimator to the true underlying distribution function.

Keywords: Self-consistency algorithm; Uniqueness, Convergence.

1 Introduction

Three data collection schemes have been referred to as *interval censored.* Following Peto (1973) we use this term only to refer to the following situation. For each individual i there is a sequence of inspection times $t_{i,1}$, $t_{i,2}$, ... The exact failure time x_i of the individual is not observed. All that is known is which inspection times immediately preceded and followed the failure (the j such that $t_{i,j-1} < x_i < t_{i,j}$). Such data have been considered by Peto (1973), Turnbull(1976), and Finkelstein (1986), among others. A generalization of this situation has been considered by De Gruttola and Lagakos (1989), but they refer to it as doubly-censored data. Interval censored data, as we have defined it, differs substantially from grouped data (Heitjan, 1989) and the doubly-censored data of Chang and Yang (1987).

2 The Likelihood

Suppose that survival times, X, arise from a distribution F_0, that each individual has a possibly infinite sequence of inspection times arising from some stochastic process Q, and that the inspection times and failure times are independent (so the censoring is noninformative). Also suppose that no time point occurs with positive probability under the inspection time process. This assumption is made to ensure that failures cannot coincide with inspection times. The observed data consist of the last inspection time prior to failure and the first inspection time after failure for each individual, i. e., the data are $\{A_i\}_{i=1}^{n}$ where $A_i = (L_i, R_i)$ is the open interval known to contain the unobserved failure time.

These assumptions ensure that the probabilities of inspection times do not involve any of the parameters of interest and hence we may consider the likelihood conditional upon the observed intervals,

$$L = \prod_{i=1}^{n} \{F_0(R_i-) - F_0(L_i)\}.$$

Let $\{s_j\}_{j=0}^{m}$ denote the unique ordered elements of $\{0, \{L_i\}_{i=1}^{n}, \{R_i\}_{i=1}^{n}\}$. Then as noted in Peto (1973), Turnbull (1976), and Finkelstein (1986) the likelihood depends on F_0 only through the values $\{F_0(s_j)\}_{j=1}^{m}$ and not on how F changes between the

s_j. Let α_{ij} be the indicator of the event $(s_{j-1}, s_j) \in A_i$ and $p_j = F_0(s_j-) - F_0(s_{j-1})$ then the likelihood can be written

$$L = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha_{ij} p_j \right]$$

and the log likelihood as

$$l(\mathbf{p}) = \sum_{i=1}^{n} \log(\sum_{j=1}^{m} \alpha_{ij} p_j).$$

Also let

$$d_k = \frac{\partial l}{\partial p_k} = \sum_{i=1}^{n} \frac{\alpha_{ik}}{\eta_i},$$

where

$$\eta_i = \sum_{j=1}^{m} \alpha_{ij} p_j.$$

The terms η_i correspond to the sum of probabilities associated with the i^{th} individual and hence d_k is the sum of $1/\eta_i$ for all individuals whose intervals, A_i, intersect the interval (s_{k-1}, s_k).

2.1 The Kuhn-Tucker Conditions

To find the maximum likelihood estimate of the vector $\mathbf{p}$ we maximize $l(\mathbf{p})$ with respect to $\mathbf{p}$ subject to the constraints

$$1 - \sum_{j=1}^{m} p_j = 0 \tag{1}$$

$$p_j \geq 0, \qquad j = 1, \ldots, m \tag{2}$$

For a concave programming problem with linear constraints, the Kuhn-Tucker conditions are necessary and sufficient for optimality (Rockafellar, 1970, Theorem 28.1 and Corollary 28.2.2). A point $\hat{\mathbf{p}}$ is an MLE if and only if there exist Lagrange multipliers μ_j, $j = 0, \ldots, m$ such that the Kuhn-Tucker conditions (1) through (5) hold.

$$\mu_j \cdot p_j = 0, \qquad j = 1, \ldots, m \tag{3}$$

$$\mu_j \geq 0, \qquad j = 1, \ldots, m \tag{4}$$

$$\frac{\partial}{\partial p_j} \left(l(\mathbf{p}) + \sum_{j=1}^{m} p_j(\mu_j - \mu_0) \right) = d_j + \mu_j - \mu_0 = 0, \qquad j = 1, \ldots, m, \tag{5}$$

Multiplying Equation 5 by p_j and summing yields

$$\mu_0 = \sum_j d_j p_j = \sum_{i,j} \frac{\alpha_{ij} p_j}{\eta_i} = \sum_i \frac{\eta_i}{\eta_i} = n$$

(since $\mu_j p_j = 0$ by Equation 3). If $p_j > 0$ then Equation 3 implies that $\mu_j = 0$, and Equation 5 then implies that $d_j = \mu_0 = n$. Conversely, if $p_j = 0$ then Equation 5 implies that $\mu_j \geq 0$ so $d_j = \mu_0 - \mu_j$ implies $d_j \leq n$. At a solution all of the η_i are strictly positive, since otherwise the d_k would not be finite.

We will later use another way of phrasing the Kuhn-Tucker conditions in a more descriptive terminology. For any $\mathbf{p}$ that satisfies the constraints (1) and (2), set $\mu_j = n - d_j$, if $p_j = 0$, and $\mu_j = 0$, if $p_j > 0$. Then the "complementary slackness" condition (3) is always satisfied. We call the μ_j Lagrange multipliers whether or not they satisfy the dual constraints (4). The left hand side of Equation 5), $d_j + \mu_j - n$ is called the reduced gradient, because it is the gradient with respect to the free variables. The Kuhn-Tucker conditions are satisfied if the Lagrange multipliers are nonnegative and the reduced gradient is zero.

The parameterization in terms of the p_j is often an overparameterization. Peto (1973) and Turnbull (1976) point out that p_j can be nonzero only if s_{j-1} is a left endpoint L_i for some individual i and s_j is a right endpoint R_k for some possibly different individual k. However, some of the p_j satisfying this criterion may also be zero. Determination of which of the p_j are zero is discussed in Section 4.

2.2 Uniqueness of the Maximum Likelihood Estimate

The MLE need not be unique. Turnbull (1976) gives the example where $\alpha_{ij} = \alpha_{ik}$ for all i. The maximum likelihood estimate will be unique if the log likelihood is strictly concave, that is if the Hessian $\mathbf{H}$ is strictly negative definite. Let $\mathbf{A}$ denote the n by m matrix with elements α_{ij}, then

$$\mathbf{H} = \mathbf{A}'\mathbf{D}\mathbf{A},$$

where $\mathbf{D}$ is the diagonal matrix with elements $-1/\eta_i^2$. Hence, $\mathbf{H}$ will be of full rank and the MLE will be unique if the rank of $\mathbf{A}$ is equal to m.

There may be situations in which the likelihood is concave, but not strictly concave, and the MLE is unique nevertheless. Theorem 9.3.2 of Fletcher (1987) specialized to our problem states the following. Let $\hat{\mathbf{p}}$ be a solution to the Kuhn-Tucker equations with suitable Lagrange multipliers μ. Define

$$W = \{\, \mathbf{w} \in \mathrm{I\!R}^m : w_j = 0, \text{ if } \mu_j > 0; \quad w_j \geq 0, \text{ if } \hat{p}_j = 0; \quad \textstyle\sum_j w_j = 0 \,\}.$$

Then the MLE $\hat{\mathbf{p}}$ is unique if

$$\mathbf{w}'\mathbf{H}\mathbf{w} < 0, \qquad \text{whenever } \mathbf{w} \in W \text{ and } \mathbf{w} \neq 0. \tag{6}$$

We can get a condition much easier to verify if we drop some of the constraints and verify the condition (6) with the set W replaced by

$$W' = \{\, \mathbf{w} \in \mathrm{I\!R}^m : w_j = 0, \text{ if } \mu_j > 0 \,\}.$$

Since we check a larger set, this condition implies the other and is sufficient.

From the structure of $\mathbf{H}$ this can be further simplified. Let $\mathbf{A} = (\mathbf{A}_1\ \mathbf{A}_2)$ be a partition of $\mathbf{A}$ into the columns for j such that $\mu_j > 0$ ($\mathbf{A}_1$) and the rest ($\mathbf{A}_2$). Also partition vectors $\mathbf{w} = (\mathbf{w}_1\ \mathbf{w}_2)$ in the same way. The sufficient condition involving W' can then be stated as

$$\mathbf{w}_2'\mathbf{A}_2'\mathbf{D}\mathbf{A}_2\mathbf{w}_2 \neq 0, \qquad \mathbf{w}_2 \neq 0,$$

where one direction of the inequality comes from (6) and the other from concavity. Since $\mathbf{D}$ is negative definite, this occurs if and only if $\mathbf{A}_2\mathbf{w}_2 \neq 0$, which proves the following.

Theorem 1 *A sufficient condition for uniqueness of the MLE is that the matrix* $\mathbf{A}_2$ *consisting of the columns of* $\mathbf{A}$ *corresponding to* j *such that* $\mu_j = 0$ *has rank equal to its number of columns.*

3 Consistency

Maximum likelihood estimation for interval censored data is strongly consistent. The MLE converges almost surely to the the truth (in a topology to be described presently). For simplicity we assume that $F_0(0) = 0$, and that all of the inspection times are greater than zero. We also assume that with probability one there are only a finite number of inspection times in any finite interval so that each realization of the inspection time process can be written $t = (t_0, t_1, \ldots, t_{m(t)})$ where

$$0 = t_0 < t_1 < \cdots < t_{m(t)} = +\infty$$

and $m(t)$ is either finite or ∞. (The assumption that all times are positive serves merely to avoid doubly infinite sequences here.)

The log likelihood for our problem is then

$$l(F) = \sum_{i=1}^{n} \sum_{j=1}^{m(t_i)} 1_{[t_{i,j} > x_i > t_{i,j-1}]} \log[F(t_{i,j}-) - F(t_{i,j-1})].$$

A proof of consistency requires a suitable compactification of the parameter space, which we take to be the set $\overline{\Theta}$ of all subdistribution functions with the topology of vague convergence (which is compact by Helley's selection theorem). The expectation of the log likelihood ratio, $\lambda(F) = E[l(F) - l(F_0)]$, is an upper semicontinuous, nonnegative concave function by Fatou's lemma, Jensen's inequality, and the assumption that no inspection time occurs with positive probability. So the set $C = \{\, F : \lambda(F) = 0 \,\}$ is a closed subset of $\overline{\Theta}$. The distribution functions in C cannot be distinguished by maximum likelihood. Hence, following Redner (1981), we identify all of the points in C with F_0.

Then we have the following theorem, which is proved in the Appendix A.

Theorem 2 *Under the assumptions stated above, the maximum likelihood estimate* $\{\hat{F}_n\}$ *converges strongly to the equivalence class* C *of the true distribution in the topology of vague convergence.*

This says that the sequence $\{\hat{F}_n\}$ is eventually in every neighbourhood of C.

The equivalence class C is the set of all distribution functions F such that $F(t_j) = F_0(t_j)$, $j = 0, \ldots, m(t)$, for almost all inspection time sequences $t = (t_0, t_1, \ldots, t_{m(t)})$. If the inspection time process densely samples $[0, \infty)$, the equivalence class C will contain only F_0.

4 Computation

The method proposed by Turnbull (1976), a version of the EM algorithm, is easy to implement but is known to have slow convergence. Alternative methods are the constrained Newton-Raphson method of Peto (1973) and the similar active set methods of optimization theory (Fletcher, 1987, Section 11.2). The latter are more difficult to implement but have quadratic convergence.

Another problem with Turnbull's algorithm is that there can be self-consistency points other than the MLE. These are not stationary points of the log likelihood. They are maxima relative to faces of the parameter space, but moving away from such points into the interior increases the likelihood. An example of this is the situation where $F(t)$ puts mass only on the interval $(0, 3]$. Suppose that the data are the intervals $(0, 1]$, $(1, 3]$, $(1, 3]$, $(0, 2]$, $(0, 2]$, $(2, 3]$. Then it can be verified that $p(0, 1] = 1/2$, $p(1, 2] = 0$, $p(2, 3] = 1/2$ is a self-consistent estimator while $p(0, 1] = 1/3$, $p(1, 2] = 1/3$, $p(2, 3] = 1/3$ is the maximum likelihood estimate. An examination of the Kuhn-Tucker conditions at $(1/2, 0, 1/2)$ shows that they are violated at this point.

The occurrence of self-consistency points other than the MLE is troubling for two reasons. First, continuity of the EM steps implies that the algorithm makes arbitrarily small steps near a self-consistency point so it is not possible to test for convergence by examining the sequence of iterates (or the likelihood along the sequence). Second, as will be illustrated in the next section, it is a reasonable procedure to restart the EM algorithm with very small parameter values set to zero to "polish" the parameter values. This will produce incorrect results if the zeros are incorrectly determined, since the EM iteration never changes the zeros.

Both of these problems can be cured by the simple expedient of examining the Kuhn-Tucker conditions. If they are used as the convergence test, convergence to the MLE is guaranteed. The computational effort required to check the Kuhn-Tucker conditions is minimal. All of the necessary quantities are calculated during the self-consistency iteration. Interestingly, Turnbull does derive a characterization of the MLE equivalent to the Kuhn-Tucker conditions, but he does not recommend that it be used to test for convergence of the self-consistency algorithm.

5 Example

The data in Table 1 comes from Finkelstein and Wolfe (1985). It represents the interval in which cosmetic deterioration for early breast cancer patients treated

Table 1: Intervals in which Deterioration Occurred

(45,–]	(6,10]	(0,7]	(46,–]	(46,–]	(7,16]	(17,–]	(7,14]
(37,44]	(0,8]	(4,11]	(15,–]	(11,15]	(22,–]	(46,–]	(46,–]
(25,37]	(46,–]	(26,40]	(46,–]	(27,34]	(36,44]	(46,–]	(36,48]
(37,–]	(40,–]	(17,25]	(46,–]	(11,18]	(38,–]	(5,12]	(37,–]
(0,5]	(18,–]	(24,–]	(36,–]	(5,11]	(19,35]	(17,25]	(24,–]
(32,–]	(33,–]	(19,26]	(37,–]	(34,–]	(36,–]		

with radiotherapy occurred in 46 individuals. The Kuhn-Tucker conditions indicate that there are only 14 intervals that need be considered; these intervals and the p_j associated with them are reported in Column 2 of Table 2. The matrix of α_{ij} is of full rank, hence the maximum likelihood estimate is unique.

Inspection of the probabilities indicates that several of them are very small and hence may be zero at the maximum likelihood estimate. They were set to zero and the EM algorithm applied to the resulting renormalized probability vector. The new candidate optimal point is reported in column 3 of Table 2, the reduced gradient (defined in Section 2.1) at this point is reported in column 4 and the associated Lagrange multipliers in column 5. Notice that the Kuhn-Tucker conditions are approximately satisfied at the point reported in columns 3–5 of Table 2, hence we have found the maximum likelihood estimate at a point where six of the p_j are zero.

In this problem p_2 may be set to zero without any of the η_i becoming zero. Doing this and applying the EM algorithm yields a self-consistent estimator that is not the maximum likelihood estimator as was described previously. However, an examination of the reduced gradient and the Lagrange multipliers at this point, Columns 6, 7 and 8 of Table 2, indicates that the Lagrange multiplier associated with p_2 is negative and hence the Kuhn-Tucker conditions are violated at this point. It cannot be a maximum likelihood estimate.

Table 2: Restricted Set of Intervals and the Associated Probabilities

Left	Right	Probability	Probability	Reduced Gradient	Lagrange Multiplier	Probability	Reduced Gradient	Lagrange Multiplier
4	5	0.0463	0.0463	0.0002	0	0.0583	0.0000	0
6	7	0.0335	0.0334	0.0009	0	0	0	−6.097
7	8	0.0886	0.0886	0.0003	0	0.1128	0.0000	0
11	12	0.0708	0.0708	−0.0001	0	0.0680	0.0000	0
15	16	$4.19 \cdot 10^{-19}$	0	0	24.3	0	0	24.45
17	18	$2.65 \cdot 10^{-6}$	0	0	7.65	0	0	7.091
24	25	0.0927	0.0926	0.0000	0	0.0927	0.0000	0
25	26	$1.58 \cdot 10^{-7}$	0	0	9.36	0	0	9.42
33	34	0.0817	0.0818	0.0000	0	0.0817	0.0000	0
34	35	$4.88 \cdot 10^{-8}$	0	0	10.5	0	0	10.52
36	37	0.0007	0	0	2.87	0	0	2.87
38	40	0.1174	0.1206	0.0000	0	0.1185	0.0001	0
40	44	0.0031	0	0	2.79	0	0	2.786
46	48	0.4653	0.4658	0.0000	0	0.4654	0.0000	0

A Proof of Consistency

The proof of Theorem 2 relies on the method of Wang (1985) and the idea of Redner (1981), which permit a simple proof of consistency. We first need to clarify one difference between our problem and Wang's setup. She assumes (p. 933) that the family of distributions constituting the model is dominated by a single σ-finite measure, but her method actually applies to general M-estimation. We only need to verify her assumptions for the log likelihood in our problem.

The log likelihood for a single individual in our problem is

$$l(F) = \sum_{j=1}^{m(t)} 1_{[t_j > x > t_{j-1}]} \log[F(t_j-) - F(t_{j-1})] \tag{8}$$

Note that exactly one of the indicators is nonzero so (8) is another notation for $\log[F(R-) - F(L)]$.

Wang's assumptions are as follows.

Assumption 1: There is a compactification $\overline{\Theta}$ of the parameter space Θ which is a separable metric space. Θ is the space of all distribution functions (satisfying $F(0) = 0$), and $\overline{\Theta}$ is the space of all subdistribution functions (nondecreasing, right continuous, $F(0) = 0$, and $F(\infty) \leq 1$) with the equivalence class C defined by (7) identified as one point. This is compact by the Helley selection theorem and metrized by Lévy distance if the state space $[0, \infty]$ is mapped homeomorphically to [0,1]. It is separable by Billingsley (1968, p. 239).

Assumption 2: There is a decreasing sequence of neighbourhoods V_r, $r = 1$, 2, ..., which we take to be Lévy neighbourhoods of radius $1/r$ of the equivalence class C, and for each r and F there is another parameter point, which we take to be $F_{F,r} = \frac{1}{r+1}F + \frac{r}{r+1}F_0$, such that (1) $F_{F,r}$ is in Θ (is proper) when F is, (2) $F_{F,r} \in V_r$, and (3) $l(F) - l(F_{F,r})$ is locally dominated on $\overline{\Theta}$. (1) is obvious. (2) holds because the Kolmogorov-Smirnov distance between F_0 and $F_{F,r}$ is less than or equal to $1/(r+1)$ and the Kolmogorov-Smirnov distance is dominated by the Lévy distance. For (3) we actually show global domination, that the function

$$\begin{aligned}
&\sup_{F \in \overline{\Theta}} [l(F) - l(F_{F,r})] \\
&= \sup_{F \in \overline{\Theta}} \sum_{j=1}^{m(t)} 1_{[t_j > x > t_{j-1}]} \log \frac{F(t_j-) - F(t_{j-1})}{\frac{1}{r+1}[F(t_j-) - F(t_{j-1})] + \frac{r}{r+1}[F_0(t_j-) - F_0(t_{j-1})]} \\
&\leq \sum_{j=1}^{m(t)} 1_{[t_j > x > t_{j-1}]} \log(r+1) \\
&\leq \log(r+1)
\end{aligned}$$

has finite expectation. As an aside,. this shows the advantage of Wang's method. Classical methods require that $l(F) - l(F_0)$ be locally dominated, which is not true.

Assumption 3: For any $F \notin C$ the expectation of $l(F) - l(F_{F,r})$ is strictly less than zero. The proof is similar to Lemma 4.4 in Wang (1985).

Assumption 4: $l(F) - l(F_{F,r})$ is lower semicontinuous at each $F \in \overline{\Theta}$ except for a null set of points which does not depend on F. Actually no null set is necessary. By the portmanteau theorem (Billingsley, 1968, p. 11) if F_n converges vaguely to F then $\liminf_n F_n(t_j-) - F_n(t_{j-1}) \geq F(t_j-) - F(t_{j-1})$. This, together with the monotonicity of $u \mapsto \log[u/(au+b)]$ with a, b and u positive gives the result.

Assumption 5: $l(F) - l(F_{F,r})$ is upper semicontinuous at each $F \in \overline{\Theta}$ except for a null set of points which may depend on F. Here the exception set is the set of (x,t) such that no t_j is a jump of F. For such t we have by the portmanteau theorem $\lim_n F_n(t_j-) - F_n(t_{j-1}) = F(t_j-) - F(t_{j-1})$ if $F_n \to F$ vaguely.

References

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* New York: Wiley.

CHANG, M. N. & YANG, G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15**, 1536–1547.

DE GRUTTOLA, V. & LAKAGOS, S. W. (1989). Analysis of doubly-censored survival data with application to AIDS. *Biometrics* **45**, 1–11.

FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845–854.

FINKELSTEIN, D. M. & WOLFE, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933–945.

FLETCHER, R. (1987). *Practical methods of Optimization*, 2nd ed. New York: Wiley.

HEITJAN, D. F. (1989). Inference from grouped continuous data: A review. *Statistical Science* **4**, 164–183.

PETO, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* **22**, 86–91.

REDNER, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9**, 225–228.

ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton University Press.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc.* B **38**, 290–295.

WANG, J.-L. (1985). Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics. *Ann. Statist.* **13**, 932–946.