ASYMPTOTIC PROPERTIES OF THE GMLE WITH CASE 2 INTERVAL-CENSORED DATA

By Qiqing Yu[a,1] Anton Schick[a], Linxiong Li[b,2] and George Y. C. Wong[c,3]

[a] *Dept. of Mathematical Sciences, Binghamton University, NY 13902, USA*

[b] *Dept. of Mathematics, University of New Orleans, LA 70148, USA*

[c] *Strang Cancer Prevention Center, Cornell University Medical School, NY 10021, USA*

October 1996

Abstract. In case 2 interval censoring the random survival time X of interest is not directly observable, but only known to have occurred before Y, between Y and Z, or after Z, where (Y, Z) is a pair of observable inspection times such that $Y < Z$. We consider the large sample properties of the generalized maximum likelihood estimator (GMLE) of the distribution function of X with case 2 interval-censored data in which the inspection times are discrete random variables. We prove the strong consistency of the GMLE at the support points of the inspection times and establish its asymptotic normality in the case of only finite many support points.

Short Title: Case 2 Model

AMS 1991 Subject Classification: Primary 62G05; Secondary 62G20.

Key words and phrases: Asymptotic normality; consistency; generalized maximum likelihood estimate; self-consistent algorithm.

[1]Partially supported by NSF Grant DMS-9402561 and DAMD17-94-J-4332.
[2]Partially supported by LEQSF Grant 357-70-4107.
[3]Partially supported by DAMD17-94-J-4332.

1. Introduction

In many biomedical studies, the random survival time X of interest is never observed and is only known to lie before an inspection time Y, between two consecutive inspection times Y and Z, or after the inspection time Z. This censoring scheme is referred to as case 2 interval censoring. Examples can be found in cancer studies (Finkelstein and Wolfe (1985)) and AIDS studies (Becker and Belbye (1991); Aragon and Eberly (1992)). We assume throughout that X and (Y, Z) are independent and that $Y < Z$ with probability one. We denote the distribution function of X by F_0 and the joint distribution function of (Y, Z) by G. The available data for the case 2 interval-censorship model are thus

$$(Y_j, Z_j, I[X_j \leq Y_j], I[Y_j < X_j \leq Z_j]), \quad j = 1, \ldots, n,$$

where $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ are independent copies of (X, Y, Z) and $I[A]$ is the indicator of the set A.

Groeneboom and Wellner (1992) considered the case 2 interval-censorship model with continuous F_0 and absolutely continuous G. They proposed an iterative convex minorant algorithm to calculate the GMLE and proved the uniform strong consistency of the GMLE. They showed that the estimator of F_0 obtained at the first step of the iterative convex minorant algorithm converges to F_0 at the $(n \log n)^{1/3}$ rate and that its asymptotic distribution is not normal. The asymptotic distribution of the GMLE remains unresolved. There are other approaches to derive the GMLE. They include Peto's (1973) Newton-Raphson algorithm and Turnbull's (1976) self-consistent algorithm.

In this paper, we assume that F_0 is arbitrary, but (Y, Z) is discrete. This assumption is used by several authors (Becker and Melbye (1991); Finkelstein (1986)). Let

$$\mathcal{A} = \{a \in \mathbb{R} : P(Y = a) + P(Z = a) > 0\}$$

be the set of all possible values of Y or Z. We establish the strong consistency of the GMLE at each point in $\mathcal{A}$. From this we can then infer the uniform strong consistency of the GMLE if F_0 is continuous and $\mathcal{A}$ is dense in $[0, \infty)$. This is done in Section 2.

In Section 3 we consider the case of finite $\mathcal{A}$. We obtain the joint asymptotic normality of the GMLE at the usual $\sqrt{n}$ rate for the points in $\mathcal{A}$ and present a consistent estimator of its asymptotic variance.

2. The consistency of the GMLE

Let $\mathcal{B}$ denote the set of all pairs (a, b) such that $g(a, b) = P(Y = a, Z = b) > 0$. In other words, $\mathcal{B}$ is the set of all possible values of (Y, Z). For $(a, b) \in \mathcal{B}$, let

$$N_n^-(a,b) = \frac{1}{n}\sum_{j=1}^{n} I[X_j \le a, Y_j = a, Z_j = b],$$

$$N_n^o(a,b) = \frac{1}{n}\sum_{j=1}^{n} I[a < X_j \le b, Y_j = a, Z_j = b],$$

$$N_n^+(a,b) = \frac{1}{n}\sum_{j=1}^{n} I[X_j > b, Y_j = a, Z_j = b],$$

$$N_n(a,b) = \frac{1}{n}\sum_{j=1}^{n} I[Y_j = a, Z_j = b].$$

Then the generalized likelihood is given by

$$\Lambda_n(F) = \prod_{(a,b)\in\mathcal{B}} F(a)^{nN_n^-(a,b)}[F(b) - F(a)]^{nN_n^o(a,b)}[1 - F(b)]^{nN_n^+(a,b)}$$

and the normalized generalized log-likelihood is

$$\mathcal{L}_n(F) = \sum_{(a,b)\in\mathcal{B}} \{N_n^-(a,b)\log[F(a)] + N_n^o(a,b)\log[F(b) - F(a)] + N_n^+(a,b)\log[1 - F(b)]\}.$$

Here and below we interpret $0\log 0 = 0$ and $\log 0 = -\infty$. In the above we let F range over the set $\mathcal{F}^*$ of all subdistribution functions. A function F_1 is called a subdistribution function if $F_1 = aF$ for some distribution function F and a number $a \in [0, 1]$. Note that $\Lambda_n(F)$ and $\mathcal{L}_n(F)$ depend on F only through the values of F at the points $a \in \mathcal{A}$. Thus there exists no unique maximizer of $\Lambda_n(F)$ over the set $\mathcal{F}^*$. However, there exists a unique maximizer $\hat{F}_n$ of $\Lambda_n(F)$ over the set $\mathcal{F}^*$ which satisfies $\hat{F}_n(x) = \sup\{\hat{F}_n(a) : a \le x, \sum_{j=1}^n (I[Y_j = a] + I[Z_j = a]) > 0\}$ for all $x \in \mathbb{R}$. Here we interpret the supremum of the empty set as 0. We call $\hat{F}_n$ the GMLE of F_0.

2.1. Theorem. *The GMLE $\hat{F}_n$ satisfies $\hat{F}_n(a) \to F_0(a)$ almost surely for all $a \in \mathcal{A}$.*

PROOF: Verify that

$$L(F) := E(\mathcal{L}_n(F)) = \sum_{(a,b)\in\mathcal{B}} g(a,b)h_{a,b}(F)$$

with

$$h_{a,b}(F) = F_0(a)\log[F(a)] + [F_0(b) - F_0(a)]\log[F(b) - F(a)] + [1 - F_0(b)]\log[1 - F(b)].$$

It is easy to check that the expression $h_{a,b}(F)$ is maximized by a nondecreasing function F into $[0,1]$ if and only if $F(a) = F_0(a)$ and $F(b) = F_0(b)$. Thus F_0 maximizes $L(F)$ and any other nondecreasing function into $[0,1]$ that maximizes $L(F)$ coincides with F_0 at the points in $\mathcal{A}$.

Note that $\mathcal{L}_n(F_0) = \dfrac{1}{n}\sum_{j=1}^n \psi(X_j, Y_j, Z_j)$, where ψ is the map defined by

$$\psi(x,y,z) = I[x \le y]\log(F_0(y)) + I[y < x \le z]\log(F_0(z) - F_0(y)) + I[z < x]\log(1 - F_0(z)).$$

Thus it follows from the SLLN that $\mathcal{L}_n(F_0) \to L(F_0)$ almost surely. By the definition of the GMLE, $\mathcal{L}_n(\hat{F}_n) \ge \mathcal{L}_n(F_0)$. Consequently,

$$\liminf_{n\to\infty} \mathcal{L}_n(\hat{F}_n) \ge \liminf_{n\to\infty} \mathcal{L}_n(F_0) = L(F_0) \quad \text{almost surely.}$$

Let Ω' denote the event on which $\liminf_{n\to\infty} \mathcal{L}_n(\hat{F}_n) \ge L(F_0)$ and, for each $(a,b) \in \mathcal{B}$, $N_n^-(a,b) \to F_0(a)g(a,b)$, $\sup_n N_n^-(a,b) = 0$ if $F_0(a) = 0$, $N_n^o(a,b) \to (F_0(b) - F_0(a))g(a,b)$, $\sup_n N_n^o(a,b) = 0$ if $F_0(b) = F_0(a)$, $N^+(a,b) \to (1 - F_0(b))g(a,b)$ and $\sup_n N_n^+(a,b) = 0$ if $F_0(b) = 1$. Fix an $\omega \in \Omega'$. Let the function F^* be a limit point of $\hat{F}_n(\cdot,\omega)$ in the sense that $\hat{F}_{k_n}(a,\omega) \to F^*(a)$ for all $a \in \mathcal{A}$ and for some sequence $\{k_n\}$ of positive integers tending to infinity. We now show that

$$L(F^*) \ge L(F_0).$$

Let $x_{k_n}(a,b)$ denote the value of the random variable

$$N_{k_n}^-(a,b)\log(\hat{F}_{k_n}(a)) + N_{k_n}^o(a,b)\log(\hat{F}_{k_n}(b) - \hat{F}_{k_n}(a)) + N_{k_n}^+(a,b)\log(1 - \hat{F}_{k_n}(b))$$

at the point ω. Thus, by the definition of Ω',

$$\liminf_{n\to\infty} \sum_{(a,b)\in\mathcal{B}} x_{k_n}(a,b) \geq L(F_0)$$

and

$$x_{k_n}(a,b) \to g(a,b)h_{a,b}(F^*)$$

for each $(a,b) \in \mathcal{B}$. Note also that $x_{k_n}(a,b) \leq 0$ for all $(a,b) \in \mathcal{B}$. Thus an application of Fatou's Lemma yields

$$\begin{aligned}
\limsup_{n\to\infty} \sum_{(a,b)\in\mathcal{B}} x_{k_n}(a,b) &= -\liminf_{n\to\infty} \sum_{(a,b)\in\mathcal{B}} -x_{k_n}(a,b) \\
&\leq -\sum_{(a,b)\in\mathcal{B}} \liminf_{n\to\infty} (-x_{k_n}(a,b)) \\
&= \sum_{(a,b)\in\mathcal{B}} g(a,b)h_{a,b}(F^*) \\
&= L(F^*).
\end{aligned}$$

Combining the above yields $L(F_0) \leq L(F^*)$. As F_0 maximizes L, we can conclude that $L(F^*) = L(F_0)$ and therefore $F^*(a) = F_0(a)$ for all $a \in \mathcal{A}$. Since ω was arbitrary and Ω' has probability one, we can infer the desired result. □

If $\mathcal{A}$ is a finite set, then it follows from the theorem that the GMLE is uniformly strongly consistent on $\mathcal{A}$. For arbitrary $\mathcal{A}$, the uniform strong consistency of the GMLE requires additional assumptions. The proofs of the following corollary and theorem are similar to Yu, Schick, Li and Wong (1996) and are thus omitted here.

2.2. Corollary. *Suppose that $\mathcal{A}$ is a closed set. Assume that $F_0(a-) = F_0(a)$ for every $a \in \mathcal{A}$ for which there is a sequence of points $\{a_i\}_{i\geq 1} \subset \mathcal{A}$ such that $a_i \uparrow a$. Then the GMLE is uniformly strongly consistent on $\mathcal{A}$, i.e., $\sup_{a\in\mathcal{A}} |\hat{F}_n(a) - F_0(a)| \to 0$ almost surely.*

We call a number x a *point of increase of* F_0 if either $F_0(x) < F_0(y)$ for all $y > x$ or $F_0(y) < F_0(x)$ for all $y < x$.

2.3. Theorem. *Suppose that F_0 is continuous and the closure of $\mathcal{A}$ contains the set of all points of increase of F_0. Then the GMLE is uniformly strongly consistent, i.e.,* $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)| \to 0$ *almost surely.*

3. The asymptotic normality of the GMLE

In this section we shall obtain the asymptotic normality of the GMLE under the assumption that $\mathcal{A}$ contains finitely many elements and

$$0 < F_0(a) < F_0(b) < 1 \quad \text{for all } a, b \text{ in } \mathcal{A} \text{ such that } a < b.$$

Note that under the current assumption the standard method for finite parametric models can be used.

Let $\mathcal{F}$ denote the set of all distribution functions F which satisfy $0 < F(a) < F(b) < 1$ for all a, b in $\mathcal{A}$ with $a < b$. For $F \in \mathcal{F}$ and $a \in \mathcal{A}$, let

$$\mathcal{L}_{n,a}(F) = \sum_{b:(a,b)\in\mathcal{B}} \left(\frac{N_n^-(a,b)}{F(a)} - \frac{N_n^o(a,b)}{F(b)-F(a)} \right) + \sum_{c:(c,a)\in\mathcal{B}} \left(\frac{N_n^o(c,a)}{F(a)-F(c)} - \frac{N_n^+(c,a)}{1-F(a)} \right),$$

$$\begin{aligned} \mathcal{L}_{n,a,a}(F) = &- \sum_{b:(a,b)\in\mathcal{B}} \left(\frac{N_n^-(a,b)}{F^2(a)} + \frac{N_n^o(a,b)}{(F(b)-F(a))^2} \right) \\ &- \sum_{c:(c,a)\in\mathcal{B}} \left(\frac{N_n^o(c,a)}{(F(a)-F(c))^2} + \frac{N_n^+(c,a)}{(1-F(a))^2} \right) \end{aligned}$$

and

$$\mathcal{L}_{n,a,b}(F) = \mathcal{L}_{n,b,a}(F) = \frac{N_n^o(a,b)}{(F(b)-F(a))^2}, \quad a, b \in \mathcal{A}, a < b.$$

Then

$$\mathcal{L}_{n,a}(F) = \frac{\partial \mathcal{L}_n(F)}{\partial F(a)} \quad \text{and} \quad \mathcal{L}_{n,a,b}(F) = \mathcal{L}_{n,b,a}(F) = \frac{\partial^2 \mathcal{L}_n(F)}{\partial F(a) \partial F(b)}, \quad a, b \in \mathcal{A}.$$

Let $a_1 < a_2 < \cdots < a_m$ denote the elements of $\mathcal{A}$. For $F \in \mathcal{F}$, let $\overset{\bullet}{\mathcal{L}}_n(F)$ denote the m-dimensional column vector with entries $(\overset{\bullet}{\mathcal{L}}_n(F))_i = \mathcal{L}_{n,a_i}(F)$, $i = 1, \ldots, m$, and $\overset{\bullet\bullet}{\mathcal{L}}_n(F)$ denote the $m \times m$ matrix with entries

$$(\overset{\bullet\bullet}{\mathcal{L}}_n(F))_{ij} = \mathcal{L}_{n,a_i,a_j}(F), \quad i, j = 1, \ldots, m.$$

Finally set

$$J = nE[\dot{\mathcal{L}}_n(F_0)(\dot{\mathcal{L}}_n(F_0))^T] = -E[\ddot{\mathcal{L}}_n(F_0)].$$

The matrix J is positive definite since

$$J = D + \sum_{1 \le i < j \le m} \frac{g(a_i, a_j)}{F_0(a_j) - F_0(a_i)} (e_i - e_j)(e_i - e_j)^T$$

where D is the diagonal matrix with positive diagonal elements

$$d_{ii} = \frac{P\{Y = a_i\}}{F_0(a_i)} + \frac{P\{Z = a_i\}}{1 - F_0(a_i)}, \quad i = 1, \ldots, m,$$

and $e_1, \ldots, e_m$ denote the standard basis in $\mathbb{R}^m$. It is easy to verify that

$$\ddot{\mathcal{L}}_n(\hat{F}_n) \to E[\ddot{\mathcal{L}}_n(F_0)] = -J.$$

It thus follows that on the event $\{\hat{F}_n \in \mathcal{F}\}$

$$0 = \dot{\mathcal{L}}_n(\hat{F}_n) = \dot{\mathcal{L}}_n(F_0) - J\Delta_n + o_p(\|\Delta_n\|),$$

where Δ_n is the m-dimensional column vector with entries $\hat{F}_n(a_i) - F_0(a_i)$, $i = 1, \ldots, m$. It follows from the CLT that $n^{1/2}\dot{\mathcal{L}}_n(F_0)$ is asymptotically normal with mean 0 and dispersion matrix J. This shows that $\Delta_n = J^{-1}\dot{\mathcal{L}}_n(F_0) + o_p(n^{-1/2})$. Thus we have the following result.

3.1. Theorem. *Suppose F_0 belongs to $\mathcal{F}$. Then*

$$n^{1/2} \begin{pmatrix} \hat{F}_n(a_1) - F_0(a_1) \\ \vdots \\ \hat{F}_n(a_m) - F_0(a_m) \end{pmatrix}$$

is asymptotically normal with mean 0 and dispersion matrix J^{-1}. A strongly consistent estimator of J is given by $-\ddot{\mathcal{L}}_n(\hat{F}_n)$.

4. References

Aragon, J. and Eberly, D. (1992). On convergence of convex minorant algorithms for distribution estimation with interval-censored data. *J. of Computational and Graphical Statistics*, 1, 129-140.

Becker, N.G. and Melbye, M. (1991). Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity. *Austral. J. Statist.*, 33, 125-133.

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

Finkelstein, D.M. and Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41, 933-945.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* Birkhäuser Verlag, Basel.

Peto, R. (1973). Experimental survival curve for interval-censored data. *Appl. Statist.*, 22, 86-91.

Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, 38, 290-295.

Yu, Q., Schick, A., Li, L., and Wong, G. (1996). Estimation of a survival function with case 1 interval-censored data. To appear in *Canad. J. Statist.*.