The Annals of Statistics
1996, Vol. 24, No. 2, 540–568

EFFICIENT ESTIMATION FOR THE PROPORTIONAL HAZARDS MODEL WITH INTERVAL CENSORING[1]

BY JIAN HUANG

University of Iowa

The maximum likelihood estimator (MLE) for the proportional hazards model with "case 1" interval censored data is studied. It is shown that the MLE for the regression parameter is asymptotically normal with $\sqrt{n}$ convergence rate and achieves the information bound, even though the MLE for the baseline cumulative hazard function only converges at $n^{1/3}$ rate. Estimation of the asymptotic variance matrix for the MLE of the regression parameter is also considered. To prove our main results, we also establish a general theorem showing that the MLE of the finite-dimensional parameter in a class of semiparametric models is asymptotically efficient even though the MLE of the infinite-dimensional parameter converges at a rate slower than $\sqrt{n}$. The results are illustrated by applying them to a data set from a tumorigenicity study.

1. Introduction. In many survival analysis problems, we are interested in the relationship between a failure time T and a vector of covariates Z. However, it is common that observations on T are subject to censoring. Besides the familiar right censoring, many other types of censored data also arise in practice. One of them is "case 1" interval censored data, in which it is only known whether the failure event has occurred before or after a censoring time Y. Thus the observable variable is $X = (Y, \delta, Z) \in R^+ \times \{0, 1\} \times R^d$, where $\delta = 1_{\{T \le Y\}}$ indicating whether T has occurred or not. "Case 1" interval censored data is also called current status data. Groeneboom and Wellner (1992) studied properties of the nonparametric maximum likelihood estimators of a distribution function with current status data and more general "case 2" interval censored data.

In this paper we consider Cox's (1972) proportional hazards model with current status data. The proportional hazards model is probably the most widely used regression model in survival analysis. In terms of cumulative hazard functions or survival functions, the proportional hazards model specifies:

$$\Lambda(t|z) = \exp(\theta' z)\Lambda(t) \quad \text{or} \quad \bar{F}(t|z) = \bar{F}(t)^{\exp(\theta' z)}.$$

Here $\Lambda(t)$ or $\bar{F}(t) = 1 - F(t)$ is the baseline cumulative hazard function or survival function of T, and Λ or $\bar{F}$ is assumed to be continuous.

Received May 1994; revised July 1995.
[1]Research supported in part by NSF Grant DMS-91-08409.
AMS 1991 *subject classifications*. Primary 62G05, 62E20; secondary 62G20, 62P99.
Key words and phrases. Current status data, interval censoring, information, maximum (profile) likelihood estimator, proportional hazards model, semiparametric model.

Current status data arises naturally in many applications. For example, it arises in animal tumorigenicity experiments; see, for example, Hoel and Walburg (1972), Finkelstein and Wolfe (1985) and Finkelstein (1986). The goal of such studies is to analyze the effect of a suspected carcinogen on the time T to observe a tumor. However, T cannot be observed exactly. Rather, animals die or are sacrificed at certain times, and are examined for the presence or absence of a tumor. If the tumor is irreversible and nonlethal, the observed times of death or sacrifice yield interval censored observations. Current status data also arises in demographic studies; see for example, Diamond, McDonald and Shah (1986) and Diamond and McDonald (1991). Closely related censoring schemes are of interest in studies of human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS); see Shiboski and Jewell (1992) and Jewell, Malani and Vittinghoff (1994). For a survey of regression models with interval censored data, see Huang and Wellner (1993).

Notice the difference between interval censoring and the usual right censoring. In a right censorship model, the observed data is $(\min(T, Y), 1_{\{T \le Y\}}, Z)$. There is probability $P\{T \le Y\}$ of observing the survival time exactly. But with current status data, we are not able to observe the exact value of the survival time at all, just $1_{\{T \le Y\}}$. It is therefore expected that statistical inference with interval censored data is more difficult.

There has been a tremendous amount of research on the proportional hazards model under right censoring in the last two decades; see, for example, Andersen and Gill (1982) and the recent books by Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993). However, we are not aware of any systematic treatment of the proportional hazards model under interval censoring when the baseline hazard function is completely unspecified. The treatment of Finkelstein (1986) assumes that the baseline survival function $\bar{F}$ is discrete and has finitely many known mass points, which reduces the problem to a finite-dimensional parametric estimation problem.

In the following discussion, we first define the maximum likelihood estimator of the regression parameter and the baseline cumulative hazard function in the proportional hazards model with "case 1" interval censored data. According to Groeneboom and Wellner (1992), the convergence rate of the nonparametric maximum likelihood estimator (NPMLE) of a distribution function with "case 1" interval censored data is only $n^{1/3}$, which is slower than the familiar $\sqrt{n}$ rate. Hence, for the proportional hazards model with current status data, we should not expect that the maximum likelihood estimator for the baseline cumulative hazard function converges faster than $n^{1/3}$. A natural question is: can we estimate the finite-dimensional regression parameter θ at $\sqrt{n}$ rate? If we can, how can we construct asymptotically efficient estimators for θ? We give positive answers to these questions in Section 3 by first confirming that the information for θ is positive. Then, in addition to the results on consistency and convergence rate of the maximum likelihood estimators, we state the main result that the maximum likelihood

estimator of θ is asymptotically normal and efficient with $\sqrt{n}$ convergence rate. In Section 4, we consider estimation of the covariance matrix for the estimator of θ. In Section 5, we apply the results to a data set from a tumorigenicity study of Hoel and Walburg (1972). In Section 6, we establish a theorem for maximum likelihood estimators for a class of semiparametric models. This theorem is used to prove the main result (Theorem 3.4) stated in Section 3. Proofs are put together in Section 7.

2. Maximum likelihood estimators of θ and Λ. The goal of this section is to define and characterize the maximum likelihood estimator $(\hat{\theta}_n, \hat{\Lambda}_n)$ of (θ_0, Λ_0) for a finite sample size n. Here θ_0 and Λ_0 are the "true" regression parameter and baseline cumulative hazard function. We will also denote the "true" baseline distribution function as F_0. The characterization is in terms of the score function for θ, and makes use of the monotonicity constraints, since the cumulative hazard function is increasing. The characterization also yields an algorithm for computing $(\hat{\theta}_n, \hat{\Lambda}_n)$.

Throughout the rest of the discussion, we assume that T and Y are independent given Z and that the joint distribution of (Y, Z) does not involve θ and F. So, for a single observation $X = (Y, \delta, Z)$, up to a factor not dependent on (θ, F), its probability density function is proportional to

$$p_{\theta,F}(x) = F(y|z)^{\delta}\bar{F}(y|z)^{1-\delta} = \left[1 - \bar{F}(y)^{\exp(\theta' z)}\right]^{\delta}\bar{F}(y)^{(1-\delta)\exp(\theta' z)}.$$

Hence the log-likelihood function is, up to an additive constant,

$$l(\theta, F) = \delta \log\left(1 - \bar{F}(Y)^{\exp(\theta' Z)}\right) + (1-\delta)e^{\theta' Z}\log \bar{F}(Y). \tag{2.1}$$

Let $(Y_1, 1_{\{T_1 \le Y_1\}}, Z_1), \ldots, (Y_n, 1_{\{T_n \le Y_n\}}, Z_n)$ be an i.i.d. sample distributed as (Y, δ, Z). Using $\Lambda = -\log(1-F)$, the log-likelihood function for the sample is

$$\begin{aligned} l_n(\theta, \Lambda) = {} & \sum_1^n \delta_i \log\left(1 - \exp\left(-\Lambda(Y_i)\exp(\theta' Z_i)\right)\right) \\ & -(1-\delta_i)\exp(\theta' Z_i)\Lambda(Y_i). \end{aligned} \tag{2.2}$$

The main advantage of this parametrization is that the log-likelihood function is concave with respect to the cumulative hazard function Λ.

Let $Y_{(1)}, \ldots, Y_{(n)}$ be the order statistics of $Y_1, \ldots, Y_n$; that is, $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$. Let $\delta_{(i)}, Z_{(i)}$ correspond to $Y_{(i)}$; that is, if $Y_{(i)} = Y_j$, then $\delta_{(i)} = 1_{\{T_j \le Y_j\}}$ and $Z_{(i)} = Z_j$. Let $\Lambda_{(i)} = \Lambda(Y_{(i)})$. Since only the values of Λ at $Y_{(i)}$'s matter in the log-likelihood function, to avoid ambiguity, we will take the maximum likelihood estimator $\hat{\Lambda}_n$ of Λ_0 as the right-continuous increasing step function with jump points at $Y_{(i)}$ and values $\hat{\Lambda}_n(Y_{(i)})$, $i = 1, \ldots, n$, where $Y_{(0)} = 0$ and $\hat{\Lambda}_n(0) = 0$. That is, $\hat{\Lambda}_n(y) = 0$ for $0 \le y < Y_{(1)}$, $\hat{\Lambda}_n(y) = \hat{\Lambda}_n(Y_{(i)})$ for $Y_{(i)} \le y < Y_{(i+1)}$, $i = 1, \ldots, n-1$. However, we do not specify $\hat{\Lambda}_n(y)$ for $y > Y_{(n)}$. So $\hat{\Lambda}_n$ is completely characterized by the vector $(\hat{\Lambda}_n(Y_{(i)}),\ i = 1, \ldots, n)$. Let $\hat{\Lambda}_{(i)} = \hat{\Lambda}_n(Y_{(i)})$. Let $\Theta \subset R^d$ be the finite-dimensional parameter space of θ.

The maximum likelihood estimator of θ_0 and Λ_0 is the $\hat{\theta}_n$ and $\hat{\Lambda}_n$ corresponding to $(\hat{\Lambda}_{(1)}, \ldots, \hat{\Lambda}_{(n)})$ that maximizes

$$\phi(\theta, \tilde{x}) = \sum_{i=1}^{n} \left\{ \delta_{(i)} \log\left(1 - \exp\left(-\exp\left(\theta' Z_{(i)}\right) x_i\right)\right) - (1 - \delta_{(i)}) \exp\left(\theta' Z_{(i)}\right) x_i \right\},$$

subject to $\theta \in \Theta$ and $0 \le x_1 \le x_2 \le \cdots \le x_n$. The monotonicity constraints are imposed because a cumulative hazard function is always nondecreasing.

Computationally, this is a maximization problem subject to linear constraints on the part of $\tilde{x}$; it can be solved by a number of constraint optimization softwares, such as NPSOL of Gill, Murray, Saunders and Wright (1986). However, the following two-step approach is very helpful in understanding the properties of $(\hat{\theta}_n, \hat{\Lambda}_n)$. This approach is exactly the maximum profile likelihood function procedure. We found it sufficient for moderate sample size and low-dimensional θ. The computation of the example of Section 5 is carried out via this approach.

(a) For any θ fixed, construct a function $\Lambda_n(\cdot; \theta)$ by maximizing $l_n(\theta, \Lambda)$ with respect to Λ under the constraint that Λ is a right-continuous nondecreasing step function.

(b) Put $\Lambda_n(\cdot; \theta)$ back into the log-likelihood function, and maximize $l_n(\theta, \Lambda_n(\cdot; \theta))$ to obtain an estimator $\tilde{\theta}_n$ of θ_0. This estimator is called the maximum profile likelihood estimator. A natural estimator of Λ_0 is $\Lambda_n(\cdot; \tilde{\theta}_n)$.

Step (a) is concerned with maximizing a concave function over a convex cone in R^n; hence it is well defined. It is convenient to notice that for any fixed θ, if $\delta_{(1)} = 0$ or $\delta_{(n)} = 1$, then to maximize $\phi(\theta, \tilde{x})$ without violating the monotonicity constraints, we have $\Lambda_n(Y_{(1)}; \theta) = 0$ or $\Lambda_n(Y_{(n)}; \theta) = \infty$. So, without loss of generality, we may assume that $\delta_{(1)} = 1$ and $\delta_{(n)} = 0$. It is seen that maximum profile likelihood estimators are exactly the same as the maximum likelihood estimators, that is

$$\hat{\theta}_n = \tilde{\theta}_n \quad \text{and} \quad \hat{\Lambda}_n(\cdot) = \Lambda_n\left(\cdot; \tilde{\theta}_n\right). \tag{2.3}$$

Since $l_n(\hat{\theta}_n, \hat{\Lambda}_n) \ge l_n(\theta, \hat{\Lambda}_n)$ for any $\theta \in \Theta$, if θ_0 is an interior point of Θ and if $\hat{\theta}_n$ is consistent (we will prove consistency in Theorem 3.2), then $\hat{\theta}_n$ will eventually lie in the interior of Θ. So, for the purpose of studying large-sample properties for $\hat{\theta}_n$, we can write down the score equation for θ: $(\partial/\partial\theta) l_n(\theta, \Lambda)|_{\theta = \hat{\theta}_n,\ \Lambda = \hat{\Lambda}_n} = 0$, that is,

$$\sum_{i=1}^{n} \left\{ \delta_i \frac{\exp\left(-\hat{\Lambda}_n(Y_i) \exp\left(\hat{\theta}_n' Z_i\right)\right)}{1 - \exp\left(-\hat{\Lambda}_n(Y_i) \exp\left(\hat{\theta}_n' Z_i\right)\right)} - (1 - \delta_i) \right\} \times \hat{\Lambda}_n(Y_i) \exp\left(\hat{\theta}_n' Z_i\right) Z_i = 0. \tag{2.4}$$

However, because of the monotonicity constraints on $\hat{\Lambda}_n$, there are no simple score equations that can characterize $\hat{\Lambda}_n$. Notice that by (2.3), $\hat{\Lambda}_n(\cdot) = \Lambda_n(\cdot; \hat{\theta}_n)$. So to characterize $\hat{\Lambda}_n(\cdot)$ is the same as to characterize $\Lambda_n(\cdot; \hat{\theta}_n)$, which is characterized by a set of inequalities and score equations. This is along the line of Proposition 1.1 of Groeneboom and Wellner (1992).

THEOREM 2.1. *Assume that* $\delta_{(1)} = 1$, $\delta_{(n)} = 0$. *Then the maximum likelihood estimator* $(\hat{\theta}_n, \hat{\Lambda}_n)$ *satisfies* 2.4, *and*

$$\sum_{j \geq i} \left\{ \delta_{(j)} \frac{\exp\left(\hat{\theta}_n' Z_{(j)}\right) \exp\left(-\exp\left(\hat{\theta}_n' Z_{(j)}\right) \hat{\Lambda}_{(j)}\right)}{1 - \exp\left(-\exp\left(\hat{\theta}_n' Z_{(j)}\right) \hat{\Lambda}_{(j)}\right)} - (1 - \delta_{(j)}) \exp\left(\hat{\theta}_n' Z_{(j)}\right) \right\} \leq 0 \tag{2.5}$$

for $i = 1, \ldots, n$, *and*

$$\sum_{i=1}^{n} \left\{ \delta_{(i)} \frac{\exp\left(\hat{\theta}_n' Z_{(i)}\right) \exp\left(-\exp\left(\hat{\theta}_n' Z_{(i)}\right) \hat{\Lambda}_{(i)}\right)}{1 - \exp\left(-\exp\left(\hat{\theta}_n' Z_{(i)}\right) \hat{\Lambda}_{(i)}\right)} - (1 - \delta_{(i)}) \exp\left(\hat{\theta}_n' Z_{(i)}\right) \right\} \hat{\Lambda}_{(i)} = 0. \tag{2.6}$$

This theorem can be proved using Fenchel's duality theorem; see, for example, Rockafellar (1970) Theorem 31.4. A simple proof can be given completely similar to that of Proposition 1.1 of Groeneboom and Wellner (1992); hence it is omitted.

REMARK 2.1. From the above discussion, we have that for any finite sample size n, with the assumption $\delta_{(1)} = 1$ and $\delta_{(n)} = 0$, the MLE $(\hat{\theta}_n, \hat{\Lambda}_n)$ exists and is finite. Because if either $\hat{\theta}_n$ or $\hat{\Lambda}_n(Y_{(n)})$ were ∞, the log-likelihood function would be $-\infty$.

REMARK 2.2. The assumption that $\delta_{(1)} = 1$ and $\delta_{(n)} = 0$ will not affect the results later at all. Since if $\delta_{(1)} = 0$ or $\delta_{(n)} = 1$, then $\hat{\Lambda}_n(Y_{(1)}) = 0$ or $\hat{\Lambda}_n(Y_{(n)}) = \infty$; so, in the log-likelihood function $l_n(\hat{\theta}_n, \hat{\Lambda}_n)$, the terms associated with $\delta_{(1)}$ and $\delta_{(n)}$ are 0, and they will not contribute anything to $l_n(\hat{\theta}_n, \hat{\Lambda}_n)$.

REMARK 2.3. It is clear that if $(\hat{\theta}_n, \hat{\Lambda}_n)$ maximizes (2.2), let $\hat{F}_n(y) = 1 - \exp(-\hat{\Lambda}_n(y))$, that is, $\hat{F}_n$ is a piecewise subdistribution function with jump points $(Y_{(1)}, \ldots, Y_{(n)})$ and values at these jump points $(1 - \exp(-\hat{\Lambda}_{(1)}), \ldots, 1 - \exp(-\hat{\Lambda}_{(n)}))$. So we can take $(\hat{\theta}_n, \hat{F}_n)$ as the maximum likelihood estimator of (θ_0, F_0). In studying large-sample properties of the estimators, we will first prove that $(\hat{\theta}_n, \hat{F}_n)$ is consistent for (θ_0, F_0). Since, for continuous F_0, we have $1 - F_0 = \exp(-\Lambda_0)$, this implies that $(\hat{\theta}_n, \hat{\Lambda}_n)$ is consistent for (θ_0, Λ_0).

The function $\Lambda_n(\cdot; \theta)$ can be computed by using the well-known "pool-adjacent-violators" algorithm, see, for example, Robertson, Wright and Dykstra (1988). However, when the sample size n is large, the number of nonlinear equations needed to be solved is of the order at least $O(n)$. Hence

this algorithm could be too slow to compute the profile likelihood curve $l_n(\theta, \Lambda_n(\cdot;\theta))$. This motivates us to consider the following characterization of $\Lambda_n(\cdot;\theta)$ which naturally yields an iterative but more efficient algorithm.

Define the processes W_Λ, G_Λ and V_Λ by

$$W_\Lambda(y) = \int_{y' \in [0,y]} \left\{ 1_{\{t \le y'\}} \frac{\exp(-\exp(\theta' z)\Lambda(y'))}{1 - \exp(-\exp(\theta' z)\Lambda(y'))} - 1_{\{t > y'\}} \exp(\theta' z) \right\} d\mathbb{Q}_n(t, y', z),$$

$$G_\Lambda(y) = \int_{y' \in [0,y]} 1_{\{t \le y'\}} \frac{\exp(2\theta' z)\exp(-\exp(\theta' z)\Lambda(y'))}{\left[1 - \exp(-\exp(\theta' z)\Lambda(y'))\right]^2} d\mathbb{Q}_n(t, y', z)$$

and

$$V_\Lambda(y) = W_\Lambda(y) + \int_{[0,y]} \Lambda(y')\, dG_\Lambda(y'), \qquad y \ge 0,$$

where $\mathbb{Q}_n$ is the empirical measure of the (unobservable) points (T_i, Y_i, Z_i), $i = 1, \ldots, n$.

THEOREM 2.2. *For any fixed θ, suppose that $\delta_{(1)} = 1$, $\delta_{(n)} = 0$. Then $\Lambda_n(\cdot;\theta)$ maximizes $l_n(\theta, \Lambda)$ if and only if $\Lambda_n(\cdot;\theta)$ is the left derivative of the greatest convex minorant of the "self-induced" cumulative sum diagram, consisting of the points*

$$\left(G_{\Lambda_n(\cdot;\theta)}(Y_{(j)}), V_{\Lambda_n(\cdot;\theta)}(Y_{(j)})\right), \qquad j = 1, \ldots, n,$$

and the origin $(0, 0)$.

The proof of Theorem 2.2 is completely analogous to that of Proposition 1.4 of Groeneboom and Wellner (1992). This theorem gives an iterative procedure to compute $\Lambda_n(\cdot;\theta)$ for any fixed θ. It proceeds as follows. Suppose $\Lambda^{(k)}(\cdot;\theta)$ is obtained at the kth iteration; then $\Lambda^{(k+1)}(\cdot;\theta)$ is computed as the left derivative of the convex minorant of the cumulative sum diagram, consisting of the points

$$\left(G_{\Lambda^{(k)}(\cdot;\theta)}(Y_{(j)}), V_{\Lambda^{(k)}(\cdot;\theta)}(Y_{(j)})\right), \qquad j = 1, \ldots, n,$$

and the origin $(0, 0)$. This algorithm was first proposed by Groeneboom and Wellner (1992) to compute the nonparametric maximum likelihood estimator of a distribution function with "case 2" interval censored data and in a class of deconvolution problems. They provided some heuristic arguments and empirical experiences that it converges much faster than the EM algorithm. Aragón and Eberly (1992) studied the convex minorant algorithm in more detail and showed that, with some modifications, the algorithm converges. Simulation studies show that in the present situation it also converges very quickly. This algorithm is used in the example considered in Section 5.

3. Main results. In this section, we state our main results. Proofs are put together in Section 7.

First, we need the following assumptions:

(A1) The finite-dimensional parameter space Θ is a bounded subset of R^d.

(A2) (a) The covariate Z has bounded support; that is, there exists z_0 such that $|Z| \le z_0$ with probability 1. (b) For any $\theta \ne \theta_0$, the probability $P\{\theta' Z \ne \theta_0' Z\} > 0$.

(A3) $F_0(0) = 0$. Let $\tau_{F_0} = \inf\{t\colon F_0(t) = 1\}$. The support of Y is an interval $S[Y] = [l_Y, u_Y]$, and $0 \le l_Y \le u_Y < \tau_{F_0}$.

(A3)* Everything is the same as (A3) except it is now assumed that $0 < l_Y \le u_Y < \tau_{F_0}$.

(A4) The cumulative hazard function Λ_0 has strictly positive derivative on $S[Y]$, and the joint distribution function $G(y, z)$ of (Y, Z) has bounded second-order (partial) derivative with respect to y.

Assumptions (A1), (A2a), (A3) and (A3)* are imposed for consistency and rate of convergence of the estimators. In particular, (A1), (A2a) and (A3)* are needed for the entropy calculation in Lemma 3.1 which is crucial for obtaining rate of convergence and for proving asymptotic normality of $\hat{\theta}_n$. Assumption (A2b) is imposed for identifiability of θ; (A4) is useful in the proof of Theorem 3.4.

3.1. *Information calculation.* It is well known that in most parametric models and many semiparametric models (such as the Cox model with right censoring), we can estimate the finite-dimensional parameter at $\sqrt{n}$ convergence rate and asymptotically efficiently. A necessary condition is that we must have positive information. With current status data, it is not clear a priori that the information is, in fact, positive. Therefore, we first calculate the information for the regression parameter in the Cox model with current status data, and show that it is, indeed, positive under reasonable assumptions.

Denote

$$Q(y, \delta, z) = \delta \frac{\bar{F}(y|z)}{1 - \bar{F}(y|z)} - (1 - \delta) \tag{3.1}$$

and

$$O(y|z) = E\big[Q^2(Y, \delta, Z)|Y = y, Z = z\big] = \frac{\bar{F}(y|z)}{1 - \bar{F}(y|z)}. \tag{3.2}$$

THEOREM 3.1. *Suppose that assumptions* (A2) *and* (A3)* *are satisfied. Then*

(a) *The efficient score function for* θ *is*

$$l_\theta^*(x) = \exp(\theta' Z) Q(y, \delta, z) \Lambda(y) \left\{ z - \frac{E\big[(Z\exp(2\theta' Z)) O(Y|Z)|Y = y\big]}{E\big[(\exp(2\theta' Z)) O(Y|Z)|Y = y\big]} \right\}.$$

(b) *The information for* θ *is*

$$I(\theta) = E\big[l_\theta^*(X)\big]^{\otimes 2} = E\left\{R(Y,Z)\left[Z - \frac{E(ZR(Y,Z)|Y)}{E(R(Y,Z)|Y)}\right]^{\otimes 2}\right\}, \tag{3.3}$$

where $a^{\otimes 2} = aa'$ *for* $a \in R^d$, *and* $R(Y,Z) = \Lambda^2(Y|Z)O(Y|Z)$.

3.2. *Asymptotic properties*. As we have shown, for each fixed sample size n, $(\hat{\theta}_n, \hat{\Lambda}_n)$ is well defined. The following theorem asserts the consistency of $\hat{\theta}_n$ and the consistency of $\hat{\Lambda}_n$ on the support of Y.

THEOREM 3.2 (Consistency). *Suppose that assumptions* (A1), (A2) *and* (A3) *are satisfied. Then*

$$\hat{\theta}_n \to \theta_0 \quad a.s.,$$

and if $y \in S[Y]$ *is a continuity point of* F_0,

$$\hat{\Lambda}_n(y) \to \Lambda_0(y) \quad a.s.$$

Moreover, if F_0 *is continuous, then*

$$\sup_{y \in S[Y]} \big|\hat{\Lambda}_n(y) - \Lambda_0(y)\big| \to 0 \quad a.s.$$

We now consider the convergence rate for $(\hat{\theta}_n, \hat{\Lambda}_n)$ under an appropriate norm defined later. After consistency of $\hat{\theta}_n$ is established, we can focus our attention on a neighborhood of θ_0. For any $\eta > 0$, let $B(\theta_0, \eta)$ be the ball centered at θ_0 and with radius η. If θ_0 is on the boundary of Θ, then take $B(\theta_0, \eta)$ to be $B(\theta_0, \eta) \cap \Theta$. In this way, we always have $B(\theta_0, \eta) \subset \Theta$. In the following, we will assume that the support $S[Y]$ of the censoring variable Y is finite and is strictly contained in the support of F_0, the distribution of the failure time T, so Λ_0 is bounded away from 0 and ∞ on $S[Y]$. Since we have proved that $\hat{\Lambda}_n$ converges on $S[Y]$, we may restrict $\hat{\Lambda}_n$ to the following class of functions:

$$\begin{aligned}\Phi = \{\Lambda\colon \Lambda \text{ is increasing and } 0 < 1/M \le \Lambda(y) \le M < \infty \\ \text{for all } y \in S[Y]\},\end{aligned} \tag{3.4}$$

where M is a large positive constant. Restrict the class of log-likelihood functions $l(\theta, \Lambda)$ defined by (2.1) to

$$\mathscr{H} = \{l(\theta, \Lambda)\colon \theta \in B(\theta_0, \eta), \Lambda \in \Phi\}. \tag{3.5}$$

We will apply Theorem 3.2.1 of Van der Vaart and Wellner (1996), which shows that the convergence rate is determined by the smoothness of the model and the modulus of continuity of the objective function (which is the log-likelihood function in the case of maximum likelihood estimation) over the parameter space.

For any probability measure Q, define $L_2(Q) = \{f\colon \int f^2\, dQ < \infty\}$. Let $\|\cdot\|_2$ be the usual L_2 norm, that is, $\|f\|_2 = (\int f^2\, dQ)^{1/2}$. For any subclass $\mathscr{F}$ of

$L_2(Q)$, define the bracketing number $N_{[\,]}(\varepsilon, \mathscr{F}, L_2(Q)) = \min\{m$: there exist $f_1^L, f_1^U, \ldots, f_m^L, f_m^U$ such that for each $f \in \mathscr{F}$, $f_i^L \le f \le f_i^U$ for some i, and $\|f_i^U - f_i^L\|_2 \le \varepsilon\}$.

LEMMA 3.1. *Let $\mathscr{H}$ be defined by* (3.5), *and suppose that Z has bounded support. Then there exists a constant $C > 0$ such that*

$$\sup_Q N_{[\,]}(\varepsilon, \mathscr{H}, L_2(Q)) \le C(1/\varepsilon^d)e^{1/\varepsilon} \quad \text{for all } \varepsilon > 0,$$

where d is the dimension of θ. Hence, for ε small enough, we have

$$\sup_Q \log N_{[\,]}(\varepsilon, \mathscr{H}, L_2(Q)) \le C(1/\varepsilon).$$

Here Q runs through the class of all probability measures.

Define the distance d on $R^d \times \Phi$ as follows:

$$d((\theta_1, \Lambda_1), (\theta_2, \Lambda_2)) = |\theta_1 - \theta_2| + \|\Lambda_1 - \Lambda_2\|_2,$$

where $|\theta - \theta_0|$ is the Euclidean distance in R^d,

$$\|\Lambda_1 - \Lambda_2\|_2 = \left[\int (\Lambda_1(y) - \Lambda_2(y))^2 \, dQ_Y(y)\right]^{1/2}$$

and Q_Y is the marginal probability measure of the censoring variable Y. Applying Lemma 3.1 and Theorem A.1 and Lemma A.1 in the Appendix, we can prove the following result.

THEOREM 3.3 (Rate of convergence). *Suppose that assumptions* (A1), (A2) *and* (A3)* *are satisfied. Then*

$$d\left((\hat{\theta}_n, \hat{\Lambda}_n), (\theta_0, \Lambda_0)\right) = O_p(n^{-1/3}).$$

The overall rate of convergence is dominated by $\hat{\Lambda}_n$. This rate agrees with the convergence rate of the NPMLE of a distribution function studied by Groeneboom and Wellner (1992). It is shown in the next theorem that the convergence rate of $\hat{\theta}_n$ can be refined to achieve $\sqrt{n}$.

THEOREM 3.4 (Asymptotic normality and efficiency). *Suppose that θ_0 is an interior point of Θ and that assumptions* (A1), (A2), (A3)* *and* (A4) *are satisfied. Then*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = I(\theta_0)^{-1} \sqrt{n} P_n l^*_{\theta_0}(x) + o_p(1) \to_d N\left(0, I(\theta_0)^{-1}\right),$$

where P_n is the empirical measure of (δ_i, Y_i, Z_i), $i = 1, \ldots, n$, $l^*_{\theta_0}(x)$ is the efficient score defined in Theorem 3.1 *and $I(\theta_0)$ is the information.*

Since $\hat{\theta}_n$ is asymptotically linear with efficient influence function, and the model (the likelihood function) is sufficiently smooth (Hellinger differentiable)

with respect to (θ, Λ), it is asymptotically efficient in the sense that any regular estimator has asymptotic variance matrix no less than that of $\hat{\theta}_n$. We do not go into the details here, but refer the reader to Van der Vaart (1991) and Bickel, Klaassen, Ritov and Wellner (1993), Chapter 3.

4. Estimation of the information matrix $I(\theta_0)$. We now consider estimation of the information matrix $I(\theta_0)$. By Theorem 3.4, $I(\theta_0)^{-1}/n$ is the asymptotic variance–covariance matrix for $\hat{\theta}_n$.

Recall in the expression for $I(\theta_0)$, $R(y, z)$ is defined to be

$$R(y, z) = \Lambda^2(y|z)O(y|z) = \exp(2z'\theta_0)\Lambda_0^2(y)\frac{\exp(-\Lambda_0(y)\exp(z'\theta_0))}{1 - \exp(-\Lambda_0(y)\exp(z'\theta_0))}.$$

Denote

$$\hat{R}_n(y, z) = \exp\bigl(2z'\hat{\theta}_n\bigr)\hat{\Lambda}_n^2(y)\frac{\exp\bigl(-\hat{\Lambda}_n(y)\exp\bigl(z'\hat{\theta}_n\bigr)\bigr)}{1 - \exp\bigl(-\hat{\Lambda}_n(y)\exp\bigl(z'\hat{\theta}_n\bigr)\bigr)}. \tag{4.1}$$

Let

$$\mu_1(y) = E(R(y, Z)|Y = y) \quad \text{and} \quad \mu_2(y) = E(ZR(y, Z)|Y = y).$$

Then, when we have obtained reasonable estimators $\mu_{1n}(y)$ and $\mu_{2n}(y)$ for $\mu_1(y)$ and $\mu_2(y)$, we can estimate $I(\theta_0)$ by

$$\hat{I}_n\bigl(\hat{\theta}_n\bigr) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{R}_n(Y_i, Z_i)\left[Z_i - \frac{\mu_{2n}(Y_i)}{\mu_{1n}(Y_i)}\right]^{\otimes 2}\right\}. \tag{4.2}$$

The hard word is to estimate $\mu_1(y)$ and $\mu_2(y)$. Also, θ_0 and the rest of R need to be estimated. When Z is a continuous covariate vector, $\mu_1(y) \equiv E(R(y, Z)|Y = y)$ can be approximated by $E(\hat{R}_n(y, Z)|Y = y)$. Then we can estimate $E(\hat{R}_n(y, Z)|Y = y)$ by nonparametric regression approaches; see, for example, Stone (1977).

When Z is a categorical covariate, the above nonparametric smoothing procedure does not work well because of the discrete nature of the values of $\hat{R}_n(y, z)$. Here we consider the simplest case when Z is a dichotomous variable indicating two treatment groups; that is, Z only takes values 0 or 1 with $P\{Z = 1\} = \gamma$ and $P\{Z = 0\} = 1 - \gamma$. Thus $E[ZR(Z, Y)|Y = y] = R(1, y)P\{Z = 1|Y = y\} = R(1, y)f_1(y)\gamma/f(y)$, and

$$\begin{aligned} E\bigl[R(Z, Y)|Y = y\bigr] &= R(1, y)P\{Z = 1|Y = y\} + R(0, y)P\{Z = 0|Y = y\} \\ &= \frac{R(1, y)f_1(y)\gamma}{f(y)} + \frac{R(0, y)f_0(y)(1 - \gamma)}{f(y)}, \end{aligned}$$

where $f_1(y)$ is the conditional density of Y given $Z = 1$, $f_0(y)$ is the conditional density of Y given $Z = 0$ and $f(y)$ is the marginal density of Y. Notice that we only need to estimate the ratio of the above two conditional expectations. $f(y)$ will cancel when we take the ratio of the two conditional expectations.

First we can estimate γ by the total number of subjects in the treatment group with $Z = 1$ divided by the sample size. Let $\hat{f}_{1n}(y)$ be a kernel density estimator of $f_1(y)$ and $\hat{f}_{0n}(y)$ be a kernel density estimator of $f_0(y)$. Then a natural estimator of $E[ZR(Z,Y)|Y = y]/E[R(Z,Y)|Y = y]$ is

$$\hat{\mu}_n(y) = \frac{\hat{R}_n(1,y)\hat{f}_{1n}(y)\hat{\gamma}_n}{\hat{R}_n(1,y)\hat{f}_{1n}(y)\hat{\gamma}_n + \hat{R}_n(0,y)\hat{f}_{0n}(y)(1-\hat{\gamma}_n)}.$$

Here $\hat{R}_n(y,z)$ is defined in (4.1). With a proper choice of the bandwidth and kernel in estimation of $f_1(y)$ and $f_0(y)$, the above estimator is consistent; see, for example, Silverman (1986). Hence a reasonable estimator of $I(\theta_0)$ is

$$\hat{I}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{R}_n(Y_i,Z_i)(Z_i - \hat{\mu}_n(Y_i))^2\right\}. \tag{4.3}$$

We will use (4.3) to obtain an estimator of the standard error of $\hat{\theta}_n$ in an example in the next section.

In the special case when Y and Z are independent, the above nonparametric smoothing is not necessary. In this case, we have

$$I(\theta_0) = E\left\{R(Y,Z)\left[Z - \frac{E_Z(ZR(Y,Z))}{E_Z R(Y,Z)}\right]^{\otimes 2}\right\},$$

where E_Z means expectation with respect to Z. We can simply estimate $I(\theta_0)$ by

$$\hat{I}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{R}_n(Y_i,Z_i)\left[Z_i - \frac{\sum_{j=1}^{n} Z_j\hat{R}_n(Y_i,Z_j)}{\sum_{j=1}^{n}\hat{R}_n(Y_i,Z_j)}\right]^{\otimes 2}\right\}. \tag{4.4}$$

Notice that it is always true that

$$E\left\{R(Y,Z)\left[Z - \frac{E(ZR(Y,Z)|Y)}{E(R(Y,Z)|Y)}\right]^{\otimes 2}\right\}$$
$$\le E\left\{R(Y,Z)\left[Z - \frac{E_Z(ZR(Y,Z))}{E_Z R(Y,Z)}\right]^{\otimes 2}\right\}.$$

If (4.4) is used improperly, that is, when there exists dependency between Y and Z, then (4.4) will overestimate $I(\theta_0)$, and hence the variance of $\hat{\theta}_n$ will be underestimated.

5. An example from a tumorigenicity study. In this section, we apply the algorithm and the asymptotic normality Theorem 3.4 to an example from a tumorigenicity study described in Hoel and Walburg (1972). See also Finkelstein and Wolfe (1985) and Finkelstein (1986). One hundred and forty-four RFM mice are assigned to either a germ-free or a conventional environment. The purpose of this study is to compare the time from begin-

ning of the study until the time to observe a tumor. Since lung tumors are nonlethal and cannot be observed before death in RFM mice, it is appropriate to treat this data as interval censored data.

We now fit a proportional hazards model to this data set. Let the censoring indicator $\delta = 1$ if lung tumor is present, otherwise $\delta = 0$. The censoring variable Y is the age at death. The covariate $Z = 1$ for the conventional mice and $Z = 0$ for the germ-free mice. Figure 1 shows the profile likelihood function versus the regression parameter θ. It is maximized at $\hat{\theta}_n = -0.55$. So $\hat{\theta}_n = -0.55$ is the value of our maximum likelihood estimator. To estimate the standard error of $\hat{\theta}_n$, we first compare the distribution of the death times of two treatment groups. Figure 2 shows the two histograms; the curves imposed on the histograms are the kernel density estimators. It is clear that the densities of the death times of two treatment groups are different. For the conventional RFM mice, most deaths occur between 500 to 800 days. However, for the germ-free RFM mice, most deaths occur between 600 to 1000 days, and more than half of them occur between 800 to 1000 days. This

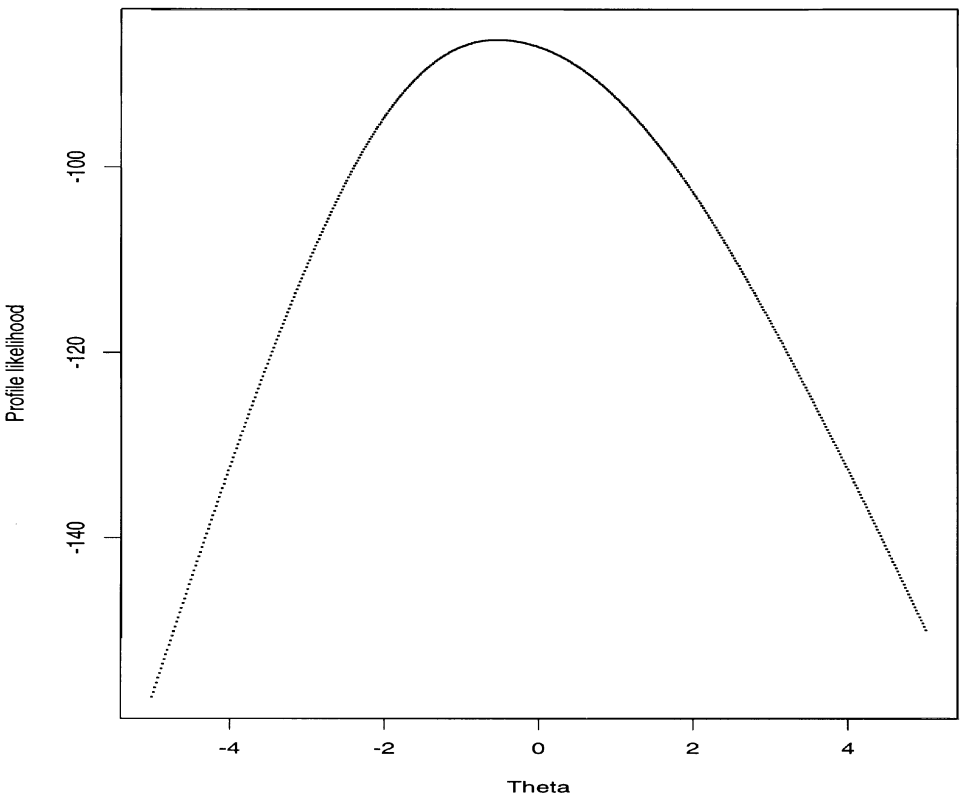

FIG. 1. *Profile likelihood function for RFM mice. The profile likelihood function is maximized at* $\hat{\theta}_n = -0.55$. *The estimated asymptotic standard error of* $\hat{\theta}_n$ *is* $\hat{\sigma}_n = 0.29$.

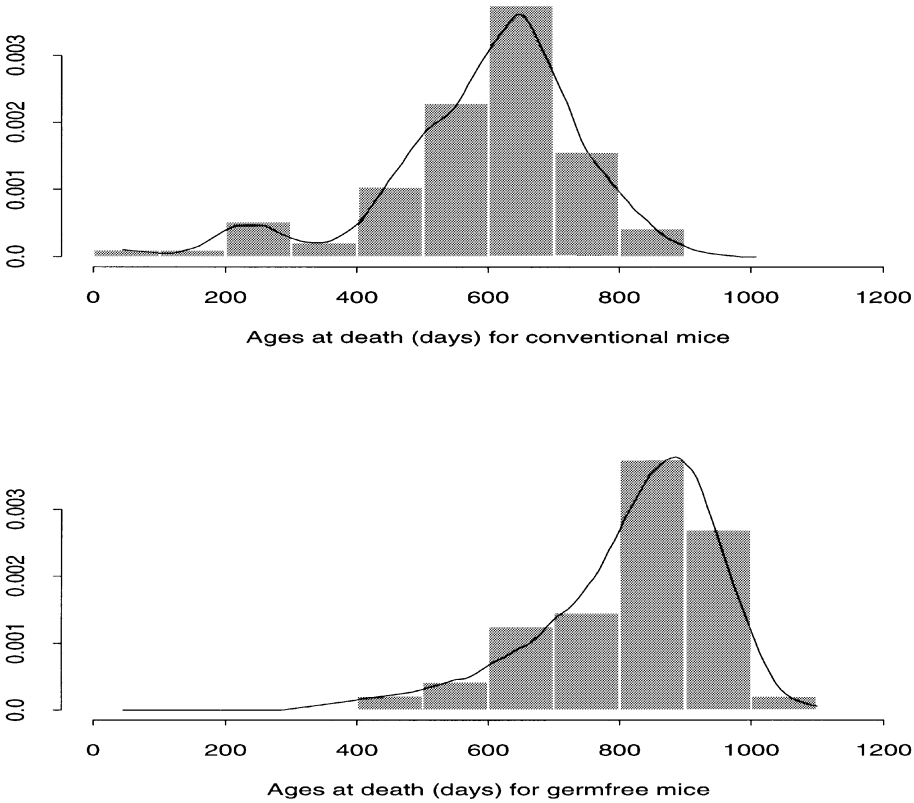

FIG. 2. *Histograms of ages at death for two groups of RFM mice. The histograms for two groups of RFM mice. The two curves are kernel density estimators using triangular kernel and bandwidth equal to* 200 *days.*

indicates a rather strong correlation between the death time and Z. So we use Theorem 3.4 and formula (4.3) to estimate the standard error of $\hat{\theta}_n$. Applying the procedure for estimating $I(\theta_0)$ using (4.3) discussed in Section 4, we obtain that the estimated standard error of $\hat{\theta}_n$ is 0.29. An approximate 95% confidence interval for θ_0 is $(-1.12, 0.01)$. So an approximate 95% confidence interval for e^{θ_0} is $(0.33, 1.01)$. The p-value for testing $\theta_0 = 0$ is 0.054. The p-value of Finkelstein's (1986) score test is 0.1. However, the approach of Finkelstein (1986) requires discretization of the data and does not give a confidence interval for θ_0.

Figure 3 shows the maximum likelihood estimator of the survival functions based on the model. The solid line is the estimator for the baseline survival function in the Cox model ($Z = 0$, germ-free mice). The dotted line is the estimator of the survival function for the conventional mice ($Z = 1$).

6. Maximum likelihood estimation in a class of semiparametric models. In this section, we consider a class of semiparametric models. In particular, we give sufficient conditions for the maximum likelihood estimator

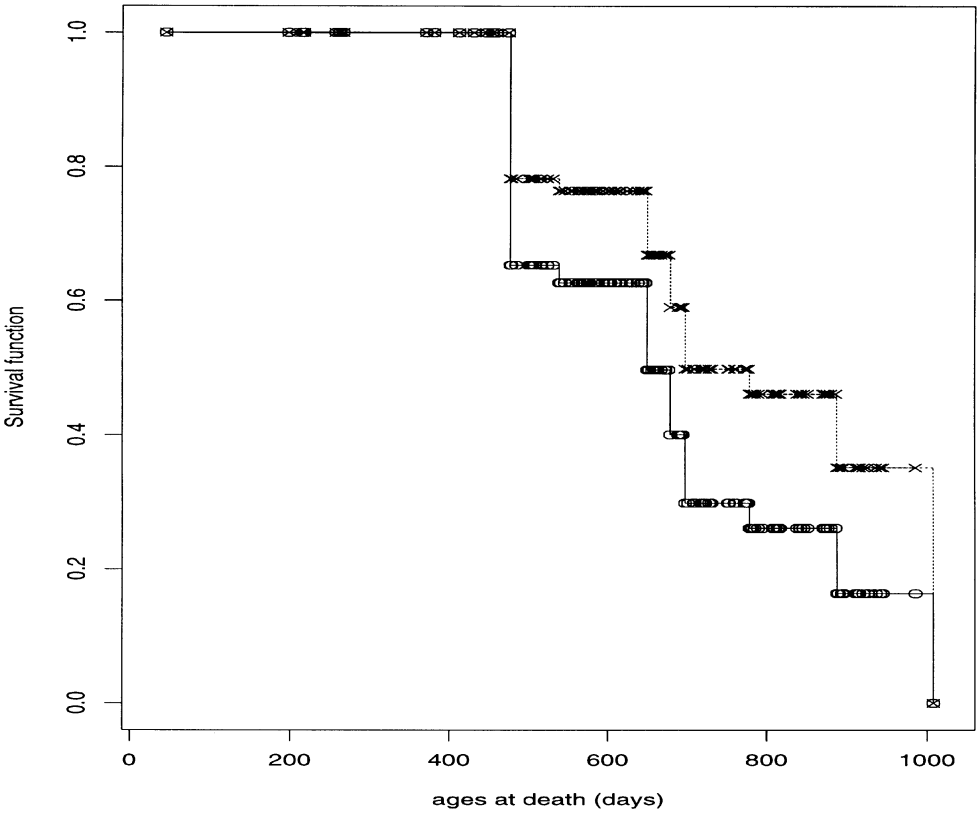

FIG. 3. *Estimated survival functions for two treatment groups. The solid line is the maximum likelihood estimator for the baseline survival function in the Cox model* ($Z = 0$, *germ-free mice*). *The dotted line is the estimator of the survival function for the conventional mice* ($Z = 1$).

of the finite-dimensional parameter in a semiparametric model to be asymptotically normal and efficient. The results are applied to prove Theorem 3.4.

Consider a semiparametric model

$$\mathscr{P} = \{p_{\theta,\phi}(x): (\theta, \phi) \in \Theta \times \Phi\},$$

where $p_{\theta,\phi}$ is a probability density function on the sample space $\mathscr{X}$, $\Theta \subset R^d$ is a finite-dimensional parameter space and Φ is an infinite-dimensional Banach space. Usually, Φ is a collection of uniformly bounded real functions.

Suppose that $X_1, \ldots, X_n$ are i.i.d. samples from a probability density function $p_{\theta_0,\phi_0}(x)$, where (θ_0, ϕ_0) is the "true value" of the parameter. Let $l_{\theta,\phi}(x) = \log p_{\theta,\phi}(x)$. The maximum likelihood estimator of (θ_0, ϕ_0) is the $(\hat{\theta}_n, \hat{\phi}_n)$ that maximizes the log-likelihood function

$$l_n(\theta, \phi) = \sum_{i=1}^{n} l_{\theta,\phi}(X_i) \quad \text{over } \Theta \times \Phi.$$

Wong and Severini (1991) showed that under some regularity conditions, even though the convergence rate of the maximum likelihood estimator of an infinite-dimensional parameter is slower than the $\sqrt{n}$ rate, a smooth functional of the maximum likelihood estimator of an infinite-dimensional parameter has $\sqrt{n}$ convergence rate and is asymptotically efficient. They also specialized their results to a class of semiparametric models, which implies that the maximum likelihood estimator of the finite-dimensional parameter is asymptotically efficient. However, their results do not seem to readily apply to the Cox model with current status data. The infinite-dimensional parameter spaces they considered are usually the spaces of smooth functions. But, for the model considered here, the natural parameter space of the infinite-dimensional parameter is the class of increasing functions, that is, the baseline cumulative hazard functions. No smoothness assumption is imposed to defined the maximum likelihood estimator. In fact, the maximum likelihood estimator of the cumulative hazard function is a right-continuous step function.

The main goal here is to propose another approach to obtain the asymptotic distribution of the maximum likelihood estimator $\hat{\theta}_n$ of the finite-dimensional parameter θ, in the presence of an infinite-dimensional nuisance parameter whose estimator may converge with a rate slower than $\sqrt{n}$.

Throughout this section, to avoid confusion with derivatives, we use a^T rather than a' to denote the transpose of vectors $a \in R^d$. Let $\theta = (\theta_1, \ldots, \theta_d)^T$. We start by defining the score functions of the log-likelihood function. For any fixed $\phi \in \Phi$, let $\{\phi(t)\colon t$ in a neighborhood of $0 \in R\}$ be a smooth curve in Φ running through ϕ at $t = 0$, that is, $\phi(0) = \phi$. Let $h = (\partial/\partial t)\phi(t)|_{t=0}$ and let H be the collection of all h defined above. Set

$$l_1(\theta, \phi; x) = \frac{\partial}{\partial \theta} l_{\theta,\phi}(x) = \left(\frac{\partial}{\partial \theta_1} l_{\theta,\phi}(x), \ldots, \frac{\partial}{\partial \theta_d} l_{\theta,\phi}(x) \right)^T,$$

$$l_2(\theta, \phi; x)[h] = \frac{\partial}{\partial t} l_{\theta,\phi(t)}(x) \Big|_{t=0}.$$

In some cases, the maximum likelihood estimator $(\hat{\theta}_n, \hat{\phi}_n)$ satisfies the following estimating equations:

$$S_{1n}\left(\hat{\theta}_n, \hat{\phi}_n\right) \equiv P_n l_1\left(\hat{\theta}_n, \hat{\phi}_n; X\right) = 0$$

and

$$S_{2n}\left(\hat{\theta}_n, \hat{\phi}_n\right)[h] \equiv P_n l_2\left(\hat{\theta}_n, \hat{\phi}_n; X\right)[h] = 0,$$

where h runs over H, and P_n is the empirical measure of the observations $X_1, \ldots, X_n$. On the other hand, in general, the exact MLE may not exist. $(\hat{\theta}_n, \hat{\phi}_n)$ may only satisfy

$$S_{1n}\left(\hat{\theta}_n, \hat{\phi}_n\right) = o_p(n^{-1/2})$$

and

$$S_{2n}\big(\hat{\theta}_n, \hat{\phi}_n\big)[h] = o_p(n^{-1/2}) \quad \text{for some } h \in H.$$

However, this will not affect the asymptotic properties of $(\hat{\theta}_n, \hat{\phi}_n)$.

Let S_1 and S_2 be the limit versions of S_{1n} and S_{2n}; that is,

$$S_1(\theta, \phi) = Pl_1(\theta, \phi, X) \quad \text{and} \quad S_2(\theta, \phi) = Pl_2(\theta, \phi; X)[h].$$

Here $P = P_{\theta_0, \phi_0}$, and we use linear functional notation $Pf = \int f(x)\, dP(x)$. Define

$$\dot{S}_{11}(\theta, \phi) = -Pl_1(\theta, \phi; X) l_1^T(\theta, \phi, X),$$

$$\dot{S}_{12}(\theta, \phi)[h] = \dot{S}_{21}^T(\theta, \phi)[h] = -Pl_1(\theta, \phi; X) l_2(\theta, \phi; X)[h]$$

and

$$\dot{S}_{22}(\theta, \phi)[h_1, h_2] = -Pl_2(\theta, \phi; X)[h_1] l_2(\theta, \phi; X)[h_2].$$

Furthermore, for $\mathbf{h} = (h_1, \ldots, h_d)^T$, where $h_k \in H$, $k = 1, \ldots, d$, denote

$$l_2(\theta, \phi, x)[\mathbf{h}] = \big(l_2(\theta, \phi, x)[h_1], \ldots, l_2(\theta, \phi, x)[h_d]\big)^T,$$

$$S_2(\theta, \phi)[\mathbf{h}] = Pl_2(\theta, \phi, x)[\mathbf{h}], \qquad S_{2n}(\theta, \phi)[\mathbf{h}] = P_n l_2(\theta, \phi, x)[\mathbf{h}],$$

$$\dot{S}_{12}(\theta, \phi)[\mathbf{h}] = \dot{S}_{21}^T(\theta, \phi)[\mathbf{h}] = -Pl_1(\theta, \phi; X) l_2^T(\theta, \phi; X)[\mathbf{h}]$$

and

$$\dot{S}_{22}(\theta, \phi)[\mathbf{h}, h] = -Pl_2(\theta, \phi; X)[\mathbf{h}] l_2(\theta, \phi; X)[h].$$

Let the infinite-dimensional space $\Phi - \phi_0 = \{\phi - \phi_0 \colon \phi \in \Phi\}$ be endowed with a seminorm $\|\cdot\|$. In applications, $\|\cdot\|$ can be chosen to be the most convenient one. Here are the assumptions needed for the result in this section.

CONDITION 1 (Consistency and rate of convergence). We have

$$|\hat{\theta}_n - \theta_0| = o_p(1) \quad \text{and} \quad \|\hat{\phi}_n - \phi_0\| = O_p(n^{-\beta}) \quad \text{for some } \beta > 0.$$

CONDITION 2 (Positive information). There is an $\mathbf{h}^* = (h_1^*, \ldots, h_d^*)^T$, where $h_k^* \in H$, $k = 1, \ldots, d$, such that

$$\dot{S}_{12}(\theta_0, \phi_0)[h] - \dot{S}_{22}(\theta_0, \phi_0)[\mathbf{h}^*, h] = 0$$

for all $h \in H$. Furthermore, $\dot{S}_{11}(\theta_0, \phi_0) - \dot{S}_{21}(\theta_0, \phi_0)[\mathbf{h}^*]$ is nonsingular.

This condition is closely related to the information calculation for a semiparametric model. We have

$$\dot{S}_{12}(\theta_0, \phi_0)[h] - \dot{S}_{22}(\theta_0, \phi_0)[\mathbf{h}^*, h] = -P\{(l_1(\theta_0, \phi_0; X) - l_2(\theta_0, \phi_0; X)[\mathbf{h}^*]) l_2(\theta_0, \phi_0; X)[h]\}.$$

So for $k = 1, \ldots, d$, if $h_k^* \in H$ and if $l_2(\theta_0, \phi_0; x)[h_k^*]$ is the projection of the kth element of $l_1(\theta_0, \phi_0; x)$ in the closure of the space generated by

$\{l_2(\theta_0, \phi_0; x)[h],\ h \in H\}$ in $L_2^0(P)$, then we will have the first part of Condition 2. Here $L_2^0(P) = \{a\colon \int a\,dP = 0,\ \int a^2\,dP < \infty\}$. Denote

$$l^*(\theta, \phi; x) = l_1(\theta, \phi; x) - l_2(\theta, \phi; x)[\mathbf{h}^*]. \tag{6.1}$$

$l^*(\theta, \phi; x)$ is called the efficient score for θ. Moreover,

$$I(\theta) = -\dot{S}_{11}(\theta, \phi) + \dot{S}_{21}(\theta, \phi)[\mathbf{h}^*] = E(l^*(\theta, \phi; X))^{\otimes 2} \tag{6.2}$$

is the information for estimation of θ, and $I(\theta)^{-1}$ is the information bound.

CONDITION 3 (Stochastic equicontinuity). For any $\delta_n \downarrow 0$ and $C > 0$,

$$\sup_{|\theta-\theta_0|\le\delta_n,\ \|\phi-\phi_0\|\le Cn^{-\beta}} \left|\sqrt{n}\,(S_{1n} - S_1)(\theta, \phi) - \sqrt{n}\,(S_{1n} - S_1)(\theta_0, \phi_0)\right| = o_p(1)$$

and

$$\sup_{|\theta-\theta_0|\le\delta_n,\ \|\phi-\phi_0\|\le Cn^{-\beta}} \left|\sqrt{n}\,(S_{2n} - S_2)(\theta, \phi)[\mathbf{h}^*] - \sqrt{n}\,(S_{2n} - S_2)(\theta_0, \phi_0)[\mathbf{h}^*]\right| = o_p(1).$$

This condition can be most conveniently verified via methods developed in empirical process theory, in particular, via entropy calculations and using maximal inequalities; see, for example, Pollard (1989, 1990) and Van der Vaart and Wellner (1996).

CONDITION 4 (Smoothness of the model). For some $\alpha > 1$ satisfying $\alpha\beta > 1/2$ and for (θ, ϕ) in the neighborhood $\{(\theta, \phi)\colon |\theta - \theta_0| \le \delta_n,\ \|\phi - \phi_0\| \le Cn^{-\beta}\}$,

$$\left|S_1(\theta, \phi) - S_1(\theta_0, \phi_0) - \dot{S}_{11}(\theta_0, \phi_0)[\theta - \theta_0] - \dot{S}_{12}(\theta_0, \phi_0)[\phi - \phi_0]\right| = o(|\theta - \theta_0|) + O(\|\phi - \phi_0\|^\alpha)$$

and

$$\left|S_2(\theta, \phi)[\mathbf{h}^*] - S_2(\theta_0, \phi_0)[\mathbf{h}^*] - \left(\dot{S}_{21}(\theta_0, \phi_0)[\mathbf{h}^*]\right)(\theta - \theta_0) - \dot{S}_{22}(\theta_0, \phi_0)[\mathbf{h}^*, \phi - \phi_0]\right| = o(|\theta - \theta_0|) + O(\|\phi - \phi_0\|^\alpha).$$

Notice the interaction between the convergence rate β and the smoothness of the model indicated by α. The faster the convergence rate, the less smoothness of the model is required.

CONDITION 5. With l^* and $I(\theta_0)$ defined in (6.1) and (6.2),

$$\sqrt{n}\,P_n l^*(\theta_0, \phi_0; x) \to_d N(0, I(\theta_0)).$$

If the information is finite and greater than 0, then this condition is satisfied because $P_n l^*$ is an average of i.i.d. random variables with mean 0 and finite second moment.

THEOREM 6.1. *Suppose that*

(i) *the estimator* $(\hat{\theta}_n, \hat{\phi}_n)$ *satisfies*

$$S_{1n}\big(\hat{\theta}_n, \hat{\phi}_n\big) = o_p(n^{-1/2}) \quad \textit{and} \quad S_{2n}\big(\hat{\theta}_n, \hat{\phi}_n\big)[\mathbf{h}^*] = o_p(n^{-1/2});$$

(ii) *Conditions* 1–5 *are satisfied.*

Then

$$\sqrt{n}\,\big(\hat{\theta}_n - \theta_0\big) = I(\theta_0)^{-1}\sqrt{n}\,P_n l^*(\theta_0, \phi_0; x) + o_p(1) \to_d N\big(0, I(\theta_0)^{-1}\big).$$

PROOF. Since $S_{1n}(\hat{\theta}_n, \hat{\phi}_n) = o_p(n^{-1/2})$, $S_{2n}(\hat{\theta}_n, \hat{\phi}_n)[\mathbf{h}^*] = o_p(n^{-1/2})$, $S_1(\theta_0, \phi_0) = 0$ and $S_2(\theta_0, \phi_0)[\mathbf{h}^*] = 0$, by Conditions 1 and 3, we have

$$\sqrt{n}\,S_1\big(\hat{\theta}_n, \hat{\phi}_n\big) - \sqrt{n}\,S_{1n}(\theta_0, \phi_0) = o_p(1)$$

and

$$\sqrt{n}\,S_2\big(\hat{\theta}_n, \hat{\phi}_n\big)[\mathbf{h}^*] - \sqrt{n}\,S_{2n}(\theta_0, \phi_0)[\mathbf{h}^*] = o_p(1).$$

These two equations plus Condition 4 imply that

$$\begin{aligned}
\text{(a)}\quad & \dot{S}_{11}\big(\hat{\theta}_n - \theta_0\big) + \dot{S}_{12}\big[\hat{\phi}_n - \phi_0\big] + \Big(o\big(|\hat{\theta}_n - \theta_0|\big) + O\big(\|\hat{\phi}_n - \phi_0\|^{\alpha}\big)\Big) \\
& - S_{1n}(\theta_0, \phi_0) = o_p(n^{-1/2}), \\
\text{(b)}\quad & \dot{S}_{21}[\mathbf{h}^*]\big(\hat{\theta}_n - \theta_0\big) + \dot{S}_{22}\big[\mathbf{h}^*, \hat{\phi}_n - \phi_0\big] \\
& \qquad + \Big(o\big(|\hat{\theta}_n - \theta_0|\big) + O\big(\|\hat{\phi}_n - \phi_0\|^{\alpha}\big)\Big) \\
& \qquad - S_{2n}(\theta_0, \phi_0)[\mathbf{h}^*] = o_p(n^{-1/2}).
\end{aligned}$$

By Condition 1 and $\alpha\beta > 1/2$, $\sqrt{n}\,O(\|\hat{\phi}_n - \phi_0\|^{\alpha}) = o_p(1)$. So, by Condition 2 and (a) minus (b), we have

$$\begin{aligned}
&\big(\dot{S}_{11} - \dot{S}_{21}[\mathbf{h}^*]\big)\big(\hat{\theta}_n - \theta_0\big) + o\big(|\hat{\theta}_n - \theta_0|\big) \\
&\quad = -\big(S_{1n}(\theta_0, \phi_0) - S_{2n}(\theta_0, \phi_0)[\mathbf{h}^*]\big) + o_p(n^{-1/2}).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\sqrt{n}\,\big(\dot{S}_{11} - \dot{S}_{21}[\mathbf{h}^*]\big)\big(\hat{\theta}_n - \theta_0\big) \\
&\quad = -\sqrt{n}\,\big(S_{1n}(\theta_0, \phi_0) - S_{2n}(\theta_0, \phi_0)[\mathbf{h}^*]\big) + o_p(1).
\end{aligned}$$

Finally, since

$$S_{1n}(\theta_0, \phi_0) - S_{2n}(\theta_0, \phi_0)[\mathbf{h}^*] = P_n l^*(\theta_0, \phi_0; X)$$

and

$$\dot{S}_{11} - \dot{S}_{21}[\mathbf{h}^*] = -I(\theta_0),$$

the result follows from the second part of Condition 2 and Condition 5. □

7. Proofs.

PROOF OF THEOREM 3.1. Without loss of generality, we only prove the theorem for $Z \in R$. The general case can be proved similarly. We first compute the score function for θ and F. The score function for θ is simply the derivative of the log-likelihood with respect to θ, that is,

$$\dot{l}_\theta(x) = ze^{\theta z}\Lambda(y)\left[\delta\frac{\bar{F}(y|z)}{1-\bar{F}(y|z)} - (1-\delta)\right].$$

Now suppose $\mathscr{F}_0 = \{F_\eta, |\eta| < 1\}$ is a regular parametric subfamily of $\mathscr{F} = \{F: F \ll \mu, \mu = \text{Lebesgue measure}\}$. Set $(\partial/\partial\eta)\log f_\eta(t)|_{\eta=0} = a(t)$. Then $a \in L_2^0(F)$ and $(\partial/\partial\eta)\bar{F}_\eta(t)|_{\eta=0} = \int_t^\infty a\,dF$. The score operator for f is

$$\dot{l}_f(a)(x) = e^{\theta z}\frac{\int_y^\infty a\,dF}{\bar{F}(y)}\left[-\delta\frac{\bar{F}(y|z)}{1-\bar{F}(y|z)} + (1-\delta)\right].$$

Let $Q(y, \delta, z)$ be defined by (3.1). Then

$$\dot{l}_\theta(x) = ze^{\theta z}\Lambda(y)Q(y,\delta,z) \quad \text{and} \quad \dot{l}_f(a)(x) = -e^{\theta z}\frac{\int_y^\infty a\,dF}{\bar{F}(y)}Q(y,\delta,z).$$

With (A2) and (A3)*, $\dot{l}_\theta$ is square integrable, and for any $a \in L_2(F)$, $\dot{l}_f(a)$ is square integrable. To calculate the efficient score $\dot{l}_\theta^*$ for θ, we need to find a function a_* so that

$$\dot{l}_\theta - \dot{l}_f a_* \perp \dot{l}_f a \quad \text{for all } a \in L_2^0(F), \tag{7.1}$$

that is, $E(\dot{l}_\theta - \dot{l}_f a_*)(\dot{l}_f a) = 0$ for all $a \in L_2^0(F)$. Let $r(z) = e^z$. Some calculation yields

$$\begin{aligned} &-E\left(\dot{l}_\theta - \dot{l}_f a_*\right)\left(\dot{l}_f a\right) \\ &\quad = E_Y\left\{\frac{\int_Y^\infty a\,dF}{\bar{F}(Y)}E\left[r(2\theta Z)O(Y|Z)\left[Z\Lambda(Y) + \frac{\int_Y^\infty a_*\,dF}{\bar{F}(Y)}\right]\middle| Y\right]\right\}, \end{aligned}$$

where $O(Y|Z)$ is defined by (3.2). Let

$$E\left\{r(2\theta Z)O(Y|Z)\left[Z\Lambda(Y) + \frac{\int_Y^\infty a_*\,dF}{\bar{F}(Y)}\right]\middle| Y\right\} = 0.$$

Then

$$\Lambda(Y)E\left[r(2\theta Z)O(Y|Z)Z|Y\right] = -\frac{\int_Y^\infty a_*\,dF}{\bar{F}(Y)}E\left[r(2\theta Z)O(Y|Z)|Y\right].$$

Thus with a_* determined by

$$\int_y^\infty a_*\,dF = -\frac{\Lambda(y)\bar{F}(y)E\left[Zr(2\theta Z)O(Y|Z)|Y=y\right]}{E\left[r(2\theta Z)O(Y|Z)|Y=y\right]},$$

(7.1) holds. Notice that a_* is only determined on the support of Y. However, $\dot{l}_f a_*$ is a square integrable function with expectation 0. So the efficient score function for θ is

$$\begin{aligned}\dot{l}_\theta^*(x) &= \dot{l}_\theta(x) - \left(\dot{l}_f a_*\right)(x)\\ &= r(\theta z)Q(y,\delta,z)\Lambda(y)\left\{z - \frac{E\left[Zr(2\theta Z)O(Y|Z)|Y=y\right]}{E\left[r(2\theta Z)O(Y|Z)|Y=y\right]}\right\}.\end{aligned}$$

The information for θ is

$$I(\theta) = E\left[\dot{l}_\theta^*(X)\right]^2 = E\left\{R(Y,Z)\left[Z - \frac{E(ZR(Y,Z)|Y)}{E(R(Y,Z)|Y)}\right]^2\right\}, \tag{7.2}$$

where $R(Y,Z) = \Lambda^2(Y|Z)O(Y|Z)$. □

We now consider the proof of Theorem 3.2. First notice that $\hat{F}_n$ is only defined on the support $S[Y]$ of the distribution of the censoring variable Y. Furthermore, since $l_Y \geq 0$, we have $\hat{F}_n(0) = 0$, according to the convention we introduced in Section 2 for $\hat{\Lambda}_n(y)$, and $\hat{F}_n(y) = 1 - \exp(-\hat{\Lambda}_n(y))$ for $y \in S[Y]$. Let $\mathscr{F}$ be the set of all Borel subprobability measures on $S[Y]$. Then $\mathscr{F}$ can be equipped with the vague topology by defining that, for any sequence $F_n \in \mathscr{F}$ and $F \in \mathscr{F}$, F_n converges vaguely to F if and only if

$$\int f dF_n \to \int f dF \quad \text{for every } f \in C_0(S[Y]), \tag{7.3}$$

where $C_0(S[Y])$ is the set of all continuous functions that vanish outside a compact subset of $S[Y]$. See Bauer (1981), Chapter 7. Let the regression parameter space Θ (a bounded subset of R^d) be equipped with the usual Euclidean topology. Then the product space $\Theta \times \mathscr{F}$ can be equipped with the product of the Euclidean topology and the vague topology. For any sequence $(\theta_n, F_n) \in \Theta \times \mathscr{F}$, we say that (θ_n, F_n) converges to (θ, F) (under the above-defined product topology) if and only if $\theta_n \to \theta$ and (7.3) holds.

The proof of Theorem 3.2 is adapted from Van der Vaart and Wellner (1992), where these authors considered consistency of the NPMLE of the mixing distribution in a mixture model.

PROOF OF THEOREM 3.2. By condition (A1), Θ is a bounded subset of R^d. Without loss of generality, we can take Θ to be a compact subset, since we can always find a compact subset Θ^* of R^d such that $\Theta \subset \Theta^*$ and work with Θ^*. Throughout the proof, we will work with subprobability measures rather than subdistribution functions. In an abuse of notation, we also use F to denote the measure μ_F induced by F, so $F(x)$ should be thought as $\mu_F[0, x]$.

Let $x = (y, \delta, z)$ and let

$$f(\theta, F; x) = \delta\left(1 - \bar{F}(y)^{\exp(\theta' z)}\right) + (1-\delta)\bar{F}(y)^{\exp(\theta' z)}.$$

Let $0 < \alpha < 1$ be a fixed constant throughout the proof. Let E be the expectation under (θ_0, F_0). For any subprobability measure F and regression parameter θ, if at least one of the two is not equal to the corresponding true value F_0 or θ_0, by concavity of the function $u \to \log u$, condition (A2b) and Jensen's inequality,

$$E\left\{\log\left[1 + \alpha\left(\frac{f(\theta, F; X)}{f(\theta_0, F_0; X)} - 1\right)\right]\right\} < 0. \tag{7.4}$$

For an open ball $\mathscr{N}$ around (θ, F), define $\bar{f}(x;\mathscr{N}) = \sup_{(\theta^*, F^*)\in\mathscr{N}} f(\theta^*, F^*; x)$. Then, for a sequence of open balls $\mathscr{N}_\varepsilon$ with radius ε shrinking to (θ, F) as $\varepsilon \to 0$, we have $\bar{f}(x;\mathscr{N}_\varepsilon) \to f(\theta, F; x)$. By (7.4), for ε sufficiently small, there is an $\eta_\varepsilon > 0$ so that

$$E\log\left[1 + \alpha\left(\frac{\bar{f}(X;\mathscr{N}_\varepsilon)}{f(\theta_0, F_0; X)} - 1\right)\right] \wedge \eta_\varepsilon < 0. \tag{7.5}$$

On the other hand, since $(\hat{\theta}_n, \hat{F}_n)$ is the maximum likelihood estimator, we have

$$\sum_{i=1}^{n} \log f\left(\hat{\theta}_n, \hat{F}_n; X_i\right) \geq \sum_{i=1}^{n} \log f(\theta_0, F_0; X_i).$$

By concavity of the function $u \to \log u$, this implies

$$\sum_{i=1}^{n} \log\left[1 + \alpha\left(\frac{f\left(\hat{\theta}_n, \hat{F}_n; X_i\right)}{f(\theta_0, F_0; X_i)} - 1\right)\right] \geq 0. \tag{7.6}$$

For any vaguely open neighborhood $\mathscr{N}_0$ of the true (θ_0, F_0), its complement in $\Theta \times \mathscr{F}$ is a vaguely closed subset of a compact set, hence also vaguely compact. The open cover $\{\mathscr{N}_{(\theta, F)}, (\theta, F) \notin \mathscr{N}_0\}$ of this complement has a finite subcover $\mathscr{N}_{(\theta_1, F_1)}, \ldots, \mathscr{N}_{(\theta_k, F_m)}$. If $(\hat{\theta}_n, \hat{F}_n)$ is not in $\mathscr{N}_0$, it is in one of the subcovers. By (7.6), we have

$$\left\{\left(\hat{\theta}_n, \hat{F}_n\right) \notin \mathscr{N}_0\right\} \subset \bigcup_{k=1}^{m} \left\{\sum_{i=1}^{n} \log\left[1 + \alpha\left(\frac{\bar{f}\left(X_i;\mathscr{N}_{(\theta_k, F_k)}\right)}{f(\theta_0, F_0; X_i)} - 1\right)\right] \wedge \eta_{(\theta_k, F_k)} \geq 0\right\}.$$

The probability of each of the sets in the union is the probability that an average of uniformly bounded and independent random variables is nonnegative. However, these random variables have negative expectation by (7.5). By Hoeffding's inequality, each of the probabilities is of the order $e^{-\varepsilon n}$, where ε can be chosen equal to $2\mu^2/(\eta_0 - \log(1-\alpha))^2$. Here $\eta_0 = \max\{\eta_{(\theta_k, F_k)}: 1 \leq k \leq m\}$, and μ is any negative number that is greater than the expectation in (7.5). This is true for all $n \geq 1$. Hence

$$\sum_{n=1}^{\infty} P\left\{\left(\hat{\theta}_n, \hat{F}_n\right) \notin \mathscr{N}_0\right\} < \infty.$$

By the Borel–Cantelli lemma, it follows that, with probability 1, $(\hat{\theta}_n, \hat{F}_n) \in \mathcal{N}_0$ for all n sufficiently large. By the definition of our product topology, this implies that

$$\hat{\theta}_n \to \theta_0 \quad \text{a.s.},$$

and $\hat{F}_n$ converges vaguely to F_0 almost surely under $P_{(\theta_0, F_0)}$. In terms of distribution functions, by Theorem 4.3.1(ii) of Chung (1974), page 81, this implies that for $y \in S[Y]$, if y is a continuity point of F_0,

$$\lim_{n\to\infty} \left(\hat{F}_n(y) - \hat{F}_n(0)\right) = F_0(y) - F_0(0)$$

almost surely with respect to $P_{(\theta_0, F_0)}$. Since $F_0(0) = 0$ and by our convention $\hat{F}_n(0) = 0$, we have

$$\lim_{n\to\infty} \hat{F}_n(y) = F_0(y), \quad P_{(\theta_0, F_0)}\text{-a.s.}$$

In particular, if F_0 is continuous, this, in turn, implies

$$\lim_{n\to\infty} \sup_{y\in S[Y]} \left|\hat{F}_n(y) - F_0(y)\right| = 0$$

almost surely with respect to $P_{(\theta_0, F_0)}$. In view of Remark 2.3, the proof is complete. □

PROOF OF LEMMA 3.1. It is known [see, e.g., Van de Geer (1993)] that for the class of functions

$$\Phi = \{\Lambda\colon \Lambda \text{ is an increasing function and } 0 < 1/M \le \Lambda \le M < \infty\},$$

where M is a constant, its ε bracketing number is of the order of $m = N_{[\,]}(\varepsilon, \Phi, L_2(P)) = O(e^{1/\varepsilon})$. This means, there exists a set of functions, $\Lambda_1^L, \Lambda_1^U, \ldots, \Lambda_m^L, \Lambda_m^U$, such that, for each $\Lambda \in \Phi$, $\Lambda_i^L \le \Lambda \le \Lambda_i^U$ for some i and $\|\Lambda_i^U - \Lambda_i^L\|_2 \le \varepsilon$. We can certainly use the following set of functions: $\Lambda_i^L - \varepsilon$, $\Lambda_i^U + \varepsilon$, $i = 1, \ldots, m$. Let $\Lambda_i^{*L} = \Lambda_i^L - \varepsilon$ and $\Lambda_i^{*U} = \Lambda_i^U + \varepsilon$. Then, for any $\Lambda \in \Phi$, we have, for some i, $\Lambda_i^{*L} + \varepsilon = \Lambda_i^L \le \Lambda \le \Lambda_i^U = \Lambda_i^{*U} - \varepsilon$ and $\|\Lambda_i^{*U} - \Lambda_i^{*L}\|_2 \le 3\varepsilon$. Since Φ is uniformly bounded away from 0, we may choose ε small enough such that all the bracketing functions stay away from 0. We first show that we can choose k points $\theta_1, \ldots, \theta_k$ in $B(\theta_0, \eta)$ such that, for any $(\theta, \Lambda) \in B(\theta_0, \eta) \times \Phi$,

$$\exp(z'\theta_j)\Lambda_i^{*L}(y) \le \exp(z'\theta)\Lambda(y) \le \exp(z'\theta_j)\Lambda_i^{*U}(y) \tag{7.7}$$

for i decided above and some $1 \le j \le k$. Since Z has bounded support, we can find a constant $0 < C_1 < \infty$ such that $\exp(z'\theta) < C_1$ for all z and $\theta \in B(\theta_0, \eta)$.

$$\begin{aligned}
&\exp(z'\theta)\Lambda(y) - \exp(z'\theta_j)\Lambda_i^{*L}(y) \\
&\quad = \exp(z'\theta)\Lambda(y) - \exp(z'\theta)\Lambda_i^{*L}(y) + \exp(z'\theta)\Lambda_i^{*L}(y) - \exp(z'\theta_j)\Lambda_i^{*L}(y) \\
&\quad \ge C_1\left(\Lambda(y) - \Lambda_i^{*L}(y)\right) - C_2|\theta - \theta_j| \ge C_1\varepsilon - C_2|\theta - \theta_j|
\end{aligned}$$

and

$$\exp(z'\theta_j)\Lambda_i^{*U}(y) - \exp(z'\theta)\Lambda(y) \geq C_1\big(\Lambda_i^{*U}(y) - \Lambda(y)\big) - C_3|\theta - \theta_j| \geq C_1\varepsilon - C_3|\theta - \theta_j|.$$

Here $C_2 > 0$ and $C_3 > 0$. So if we choose $\theta_1, \ldots, \theta_k$ such that, for any $\theta \in B(\theta_0, \eta)$, $|\theta - \theta_j| \leq \min\{C_1\varepsilon/C_2, C_1\varepsilon/C_3\}$ for some $1 \leq j \leq k$, then (7.7) holds. It is well known that the minimum value of k can be on the order of $O(1/\varepsilon^d)$.

Let

$$l_{ij}^L(x) = \delta \log\Big(1 - \exp\Big(-\exp(-z'\theta_j)\Lambda_i^{*L}(y)\Big)\Big) - (1-\delta)\exp(z'\theta_j)\Lambda_i^{*U}(y),$$

$$l_{ij}^U(x) = \delta \log\Big(1 - \exp\Big(-\exp(-z'\theta_j)\Lambda_i^{*U}(y)\Big)\Big) - (1-\delta)\exp(z'\theta_j)\Lambda_i^{*L}(y).$$

Since $\log(1 - \exp(-x))$ is an increasing function, we have, for each $(\theta, \Lambda) \in B(\theta_0, \eta) \times \Phi$,

$$l_{ij}^L(x) \leq l(\theta, \Lambda; x) \leq l_{ij}^U(x)$$

for some $1 \leq i \leq m$, $1 \leq j \leq k$. Furthermore, since Z has bounded support and the functions Λ_i^{*L} and Λ_i^{*U} are bounded away from 0,

$$\big|\exp(z'\theta_j)\Lambda_i^{*U}(y) - \exp(z'\theta_j)\Lambda_i^{*L}(y)\big| \leq C_4\big|\Lambda_i^{*U}(y) - \Lambda_i^{*L}(y)\big|,$$

$$\Big|\log\Big(1 - \exp\Big(-\exp(-z'\theta_j)\Lambda_i^{*U}(y)\Big)\Big) - \log\Big(1 - \exp\Big(-\exp(-z'\theta_j)\Lambda_i^{*L}(y)\Big)\Big)\Big| \leq C_5\big|\Lambda_i^{*U}(y) - \Lambda_i^{*L}(y)\big|,$$

for some positive constants C_4 and C_5. Thus

$$\big\|l_{ij}^U(x) - l_{ij}^L(x)\big\|_2 \leq (C_4 + C_5)\big\|\Lambda_i^{*U}(y) - \Lambda_i^{*L}(y)\big\|_2 \leq 3(C_4 + C_5)\varepsilon.$$

This implies, there exist l_{ij}^L, l_{ij}^U, $i = 1, \ldots, m$, $j = 1, \ldots, k$, such that, for any $p \in \mathscr{H}$, $l_{ij}^L \leq l \leq l_{ij}^U$ for some $1 \leq i \leq m$, $1 \leq j \leq k$, and $\|l_{ij}^U - l_{ij}^L\|_2 \leq 3(C_4 + C_5)\varepsilon$. This means that the bracketing number $N_{[\ ]}(\varepsilon, \mathscr{H}, L_2(P))$ for the class $\mathscr{H}$ is of order $mk = O(\varepsilon^{-d}e^{1/\varepsilon})$. □

PROOF OF THEOREM 3.3. We apply Theorem 3.2.1 of Van der Vaart and Wellner (1996).

By the Kullback–Leibler inequality, $El(\theta, \Lambda, X)$ is maximized at $\theta = \theta_0$ and $\Lambda = \Lambda_0$. So its first derivative at θ_0 and Λ_0 is equal to 0 (this can also be easily verified directly). Since Φ is uniformly bounded away from 0 and ∞ and Z has bounded support, Taylor expansion yields that

$$El(\theta, \Lambda, X) - El(\theta_0, \Lambda_0, X) \leq -Cd^2((\theta, \Lambda), (\theta_0, \Lambda_0))$$

for some constant $C > 0$. Thus

$$\sup_{\eta/2 \leq d((\theta, \Lambda), (\theta_0, \Lambda_0)) \leq \eta} \big(El(\theta, \Lambda, X) - El(\theta_0, \Lambda_0, X)\big) \leq -\frac{C\eta^2}{4}.$$

By Lemma 3.1 and Lemma A.1 in the Appendix,

$$E^* \sup_{d((\theta,\Lambda),(\theta_0,\Lambda_0))\le\eta} \left|\sqrt{n}\,(P_n - P)(l(\theta,\Lambda,x) - l(\theta_0,\Lambda_0,x))\right| = O(1)\eta^{1/2}\left(1 + \frac{\eta^{1/2}}{\eta^2\sqrt{n}}M\right).$$

Here E^* is the outer expectation. Let

$$\phi_n(\eta) = \eta^{1/2}\left(1 + \frac{\eta^{1/2}}{\eta^2\sqrt{n}}M\right).$$

Then $\phi_n(\eta)/\eta$ is a decreasing function, and $n^{2/3}\phi_n(n^{-1/3}) = O(\sqrt{n})$ for n large. Furthermore, by Theorem 3.2, $\hat{\theta}_n$ is consistent and $\hat{\Lambda}_n$ is uniformly consistent. Hence the conditions of Theorem 3.2.1 of Van der Vaart and Wellner (1996) are satisfied. This implies

$$d\left(\left(\hat{\theta}_n, \hat{\Lambda}_n\right), (\theta_0, \Lambda_0)\right) = O_p(n^{-1/3}). \qquad \square$$

PROOF OF THEOREM 3.4. Let

$$r(y, z; \theta, \Lambda) = \frac{\exp(-\exp(\theta' z)\Lambda(y))}{1 - \exp(-\exp(\theta' z)\Lambda(y))}.$$

The score for θ is $l_1(\theta, \Lambda; x) = ze^{\theta z}\Lambda(y)(\delta r(y, z; \theta, \Lambda) - (1 - \delta))$. By (2.4),

$$S_{1n}\left(\hat{\theta}_n, \hat{\Lambda}_n\right) \equiv P_n l_1\left(\hat{\theta}_n, \hat{\Lambda}_n; x\right) = 0.$$

For any $h \in L_2(Q_Y)$ (recall that Q_Y is the marginal distribution of Y), define

$$l_2(\theta, \Lambda; x)[h] = e^{\theta z}h(y)(\delta r(y, z; \theta, \Lambda) - (1 - \delta)).$$

l_2 can be thought of as the derivative $(\partial/\partial\varepsilon)l(\theta, \Lambda + \varepsilon h, x)|_{\varepsilon=0}$. Denote

$$\mathbf{h}^*(y) = \frac{\Lambda_0(y)E\left[Z\exp(2\theta_0 Z)O(Y|Z)|Y = y\right]}{E\left[\exp(2\theta_0 Z)O(Y|Z)|Y = y\right]}$$

and

$$S_{2n}(\theta, \Lambda)[\mathbf{h}^*] = P_n l_2(\theta, \Lambda; x)[\mathbf{h}^*].$$

From the proof of Theorem 3.1, $l_1(\theta, \Lambda; x) - l_2(\theta, \Lambda; x)[\mathbf{h}^*]$ is orthogonal to $l_2(\theta, \Lambda; x)[h]$ in $L_2^0(P)$ for any h in $L_2(Q_Y)$. Let h_k^* be the kth component of $\mathbf{h}^*$, $k = 1, \ldots, d$. Since $\hat{\Lambda}_n + \varepsilon h_k^*$ is not necessarily in the space Φ, we may not have $S_{2n}(\hat{\theta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] = 0$. Instead, we prove that

$$S_{2n}\left(\hat{\theta}_n, \hat{\Lambda}_n\right)[\mathbf{h}^*] = o_p(n^{-1/2}).$$

Since Λ_0 is a strictly increasing continuous function, its inverse Λ_0^{-1} is well defined. Let $\xi_0 = \mathbf{h}^* \circ \Lambda_0^{-1}$, that is, the composition of $\mathbf{h}^*$ on the inverse of Λ_0. Then ξ_0 is well defined on the range of Λ_0. Since $\xi(\hat{\Lambda}_n(y))$ is a right-

continuous step function and has exactly the same jump points as $\hat{\Lambda}_n(y)$, by the characterization of $\hat{\Lambda}_n$,

$$P_n\Big[\xi\big(\hat{\Lambda}_n(y)\big)\exp\big(\hat{\theta}_n z\big)\Big(\delta r\big(y, z; \hat{\theta}_n, \hat{\Lambda}_n\big) - (1-\delta)\Big)\Big] = 0.$$

By assumptions (A2a), (A3) and (A4), $\mathbf{h}^*$ has bounded derivative. This and the assumption that Λ_0 has strictly positive derivative implies that ξ_0 has bounded derivative. So, noticing $\mathbf{h}^* = \mathbf{h}^* \circ \Lambda_0^{-1} \circ \Lambda_0 = \xi_0 \circ \Lambda_0$, we have

$$\begin{aligned}
&P_n\Big[\mathbf{h}^*(y)\exp\big(\hat{\theta}_n z\big)\Big(\delta r\big(y, z; \hat{\theta}_n, \hat{\Lambda}_n\big) - (1-\delta)\Big)\Big] \\
&\quad = P_n\Big[\big(\xi_0 \circ \Lambda_0(y) - \xi_0 \circ \hat{\Lambda}_n(y)\big)\exp\big(\hat{\theta}_n z\big)\Big(\delta r\big(y, z; \hat{\theta}_n, \hat{\Lambda}_n\big) - (1-\delta)\Big)\Big] \\
&\quad = (P_n - P)\Big[\big(\xi_0 \circ \Lambda_0(y) - \xi_0 \circ \hat{\Lambda}_n(y)\big) \\
&\qquad\qquad \times \exp\big(\hat{\theta}_n z\big)\Big(\delta r\big(y, z; \hat{\theta}_n, \hat{\Lambda}_n\big) - (1-\delta)\Big)\Big] \\
&\qquad + P\Big[\big(\xi_0 \circ \Lambda_0(y) - \xi_0 \circ \hat{\Lambda}_n(y)\big)\exp\big(\hat{\theta}_n z\big)\Big(\delta r\big(y, z; \hat{\theta}_n, \hat{\Lambda}_n\big) - (1-\delta)\Big)\Big].
\end{aligned}$$

We first show that the first term converges on the order of $o_p(n^{-1/2})$. Let

$$\begin{aligned}
\psi(x; \theta, \Lambda) &= \big(\xi_0 \circ \Lambda_0(y) - \xi_0 \circ \Lambda(y)\big) \\
&\quad \times e^{\theta z}\big(\delta r(y, z; \theta, \Lambda) - (1-\delta)\big).
\end{aligned} \tag{7.8}$$

For any $\eta > 0$, consider the class of functions

$$\Psi(\eta) = \{\psi(x; \theta, \Lambda): |\theta - \theta_0| + \|\Lambda - \Lambda_0\|_2 \le \eta \text{ and } \Lambda \in \Phi\}.$$

It is verified in Lemma 7.1 that $\Psi(\eta)$ is a Donsker class. This implies

$$\sup_{\psi \in \Psi(Cn^{-1/3})} (P_n - P)\psi(x; \theta, \Lambda) = o_p(n^{-1/2}).$$

This implies the first term is on the order of $o_p(n^{-1/2})$. The second term is equal to

$$\begin{aligned}
&P\Bigg[\exp\big(\hat{\theta}_n z\big)\big(\xi_0 \circ \Lambda_0(y) - \xi_0 \circ \hat{\Lambda}_n(y)\big) \\
&\qquad \times \left(\frac{\exp\big(-\exp\big(\hat{\theta}_n z\big)\hat{\Lambda}_n(y)\big) - \exp\big(-\exp(\theta_0 z)\Lambda_0(y)\big)}{1 - \exp\big(-\exp\big(\hat{\theta}_n z\big)\hat{\Lambda}_n(y)\big)}\right)\Bigg] \\
&\qquad \le C\Big(P\big[\Lambda_0(y) - \hat{\Lambda}_n(y)\big]^2\Big)^{1/2} \\
&\qquad\quad \times \Big(P\Big[\exp\big(\hat{\theta}_n z\big)\hat{\Lambda}_n(y) - \exp(\theta_0 z)\Lambda_0(y)\Big]^2\Big)^{1/2} \\
&\qquad = O_p(n^{-2/3}),
\end{aligned}$$

by the Cauchy–Schwarz inequality and Theorem 3.3. Here C is a finite constant coming from the mean value theorem and assumptions (A2), (A3) and (A4) as well as the positive lower bound of $1 - \exp(-\exp\hat{\theta}_n z)\hat{\Lambda}_n(y))$. Thus the first assumption of Theorem 6.1 is verified. Now we verify Conditions 1–5 listed there. By Theorem 3.3, Condition 1 is satisfied with $\beta = 1/3$. The information calculation asserts that Condition 2 holds. To verify Condition 3, consider the following two classes of functions:

$$\{l_1(\theta, \Lambda; x) - l_1(\theta_0, \Lambda_0; x): |\theta - \theta_0| \le \eta, \|\Lambda - \Lambda_0\| \le \eta\},$$
$$\{l_2(\theta, \Lambda; x)[\mathbf{h}^*] - l_2(\theta_0, \Lambda_0; x)[\mathbf{h}^*]: |\theta - \theta_0| \le \eta, \|\Lambda - \Lambda_0\| \le \eta\},$$

where η is near 0. It is proved in Lemma 7.1 that the entropy numbers for the above two classes are of order $1/\eta$. This implies that these two classes are Donsker, and hence Condition 3 is satisfied. Condition 4 is satisfied with $\alpha = 2$. This can be verified straightforwardly by using a Taylor expansion. So $\alpha\beta = 2 \times 1/3 > 1/2$. Condition 5 is satisfied because the information matrix $I(\theta_0)$ is finite and positive. Thus the result follows from Theorem 6.1. □

LEMMA 7.1. *For any $\eta > 0$, define the class of functions*

$$\Psi(\eta) = \{\psi(x; \theta, \Lambda): |\theta - \theta_0| + \|\Lambda - \Lambda_0\|_2 \le \eta \text{ and } \Lambda \in \Phi\},$$

where ψ is defined in (7.8) *and Φ is defined in* (3.4). *Then the L_2 covering number $N(\varepsilon, \Psi, L_2(Q))$ of Ψ:*

$$\sup_Q N(\varepsilon, \Psi, L_2(Q)) \le \text{constant} \cdot (1/\varepsilon^d)\exp(1/\varepsilon).$$

Hence, for ε close to 0, the entropy number

$$\sup_Q \log N(\varepsilon, \Psi, L_2(Q)) \le \text{constant} \cdot 1/\varepsilon.$$

Here Q runs through all probability measures. This implies that $\Psi(\eta)$ is a Donsker class.

PROOF. It was shown in Van de Geer (1993) that for the uniformly bounded class of functions Φ defined in (3.4), its L_2 covering number is on the order of $m = O(e^{1/\varepsilon})$. This means, there exists a set of functions $\Lambda_1, \ldots, \Lambda_m$, such that, for any $\Lambda \in \Phi$, $\|\Lambda - \Lambda_i\|_2 \le \varepsilon$ for some $1 \le i \le m$. Moreover, there exist $\theta_1, \ldots, \theta_k$, where $k = O(1/\varepsilon^d)$, such that, for any $\theta \in B(\theta_0, \eta)$, $|\theta - \theta_j| \le \varepsilon$ for some $1 \le j \le k$. Now define $\psi_{ij}(x) = \psi(x; \theta_j, \Lambda_i)$, where $i = 1, \ldots, m$ and $j = 1, \ldots, k$. Then, since $\Lambda \in \Phi$, Λ is uniformly bounded from 0 and ∞. Moreover, Z is bounded, and it follows that, for any $\psi \in \Psi$,

$$\|\psi - \psi_{ij}\|_2 \le \text{constant} \cdot \varepsilon$$

for the same $1 \le i \le m$ and $1 \le j \le k$ as determined above. This implies that the covering number for the class Ψ is on the order of $mk = (1/\varepsilon^d)\exp(1/\varepsilon)$. □

APPENDIX

The following theorem is Theorem 3.2.1 of Van der Vaart and Wellner (1996). This theorem establishes the relationship between the smoothness and modulus of continuity of the objective function and the rate of convergence of a maximization estimator. It is used to prove Theorem 3.3.

THEOREM A.1 (Rate of convergence). *For each n let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set Θ. Let $\theta_n \in \Theta$ (possibly random) and $0 \le \delta_n < \eta$ be arbitrary and let $\theta \to d_n(\theta, \theta_n)$ be an arbitrary map (possibly random) from Θ to $[0, \infty)$. Let C be a generic constant. Suppose that, for every n and $\delta_n < d(\theta, \theta_n) \le \eta$,*

$$\sup_{\delta/2 < d_n(\theta, \theta_n) \le \delta,\ \theta \in \Theta_n} M_n(\theta) - M_n(\theta_n) \le -\delta^2$$

and

$$E^* \sup_{\delta/2 < d_n(\theta, \theta_n) \le \delta,\ \theta \in \Theta_n} \sqrt{n}\,\left|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)\right| \le C\phi_n(\delta)$$

for function ϕ_n such that $\delta \to \phi_n(\delta)/\delta^\alpha$ is increasing on (δ_n, η) for some $\alpha < 2$. Here E^* denotes outer expectation. Let $r_n \le C\delta_n^{-1}$ satisfy

$$r_n^2 \phi(1/r_n) \le \sqrt{n} \quad \text{every } n.$$

If the sequence $\hat{\theta}_n$ satisfies $\mathbb{M}_n(\hat{\theta}_n) \ge \mathbb{M}_n(\theta_n) - O_p(r_n^{-2})$ and is consistent for θ_n, then $r_n d_n(\hat{\theta}_n, \theta_n) = O_p^*(1)$. If the displayed conditions are valid for $\eta = \infty$, then the condition that $\hat{\theta}_n$ is consistent is unnecessary.

The following lemma is Lemma 3.2.2 of Van der Vaart and Wellner (1996). Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution P and let P_n be the empirical measure of these random variables. Denote $G_n = \sqrt{n}(P_n - P)$ and $\|G_n\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} |G_n f|$ for any measurable class of functions $\mathscr{F}$. Denote

$$J_{[\,]}(\eta, \mathscr{F}, L_2(P)) = \int_0^\eta \sqrt{1 + \log N_{[\,]}(\varepsilon, \mathscr{F}, L_2(Q))}\, d\varepsilon.$$

LEMMA A.1. *Let $\mathscr{F}$ be a uniformly bounded class of measurable functions. Then*

$$E_P^* \|G_n\|_{\mathscr{F}} \le C J_{[\,]}(\eta, \mathscr{F}, L_2(P)) \left(1 + \frac{J_{[\,]}(\eta, \mathscr{F}, L_2(P))}{\eta^2 \sqrt{n}} M\right),$$

if every f in $\mathscr{F}$ satisfies $Pf^2 < \eta^2$ and $\|f\|_\infty \le M$. Here C is a finite constant not depending on n.

Acknowledgments. This paper is based on part of my Ph.D. dissertation written under the supervision of Professor Jon Wellner. I would like to thank him for his guidance and many helpful discussions on the subject of this paper. I also would like to thank Aad Van der Vaart and Jon Wellner for

making their unpublished manuscript available to me. Thanks also go to two referees and an Associate Editor for their very constructive comments and suggestions on an earlier version of this paper. Part of the writing of the first submission of this paper was done while the author was a postdoctoral fellow at the Fred Hutchinson Cancer Research Center.

REFERENCES

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

ANDERSEN, P. K., BORGAN, Ø. GILL, R. D. and KEIDING, N. (1992). *Statistical Models Based on Counting Processes*. Springer, New York.

ARAGÓN, J. and EBERLY, D. (1992). On convergence of convex minorant algorithms for distribution estimation with interval-censored data. *J. Comput. Graph. Statist.* **1** 129–140.

BAUER, H. (1981). *Probability Theory and Elements of Measure Theory*. Academic Press, New York.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

CHUNG, K. L. (1974). *A Course in Probability Theory*. Academic Press, New York.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

DIAMOND, I. D. and MCDONALD, J. W. (1991). Analysis of current status data. In *Demographic Applications of Event History Analysis* (J. Trussell, R. Hankinson and J. Tilton, eds.) 231–252. Oxford Univ. Press.

DIAMOND, I. D., MCDONALD, J. W. and SHAH, I. H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography* **23** 607–620.

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.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

GILL, P. E., MURRAY, W., SAUNDERS, M. A. and WRIGHT, M. H. (1986). User's guide for NPSOL (Version 4.0): a Fortran package for nonlinear programming. Technical Report SOL 86-2, Dept. Operations Research, Stanford Univ.

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

HOEL, D. G. and WALBURG, H. E. (1972). Statistical analysis of survival experiments. *Journal of the National Cancer Institute* **49** 361–372.

HUANG, J. and WELLNER, J. A. (1993). Regression models with interval censoring. *Proceedings of the Kolmogorov Seminar, Euler Mathematics Institute*. St. Petersburg, Russia. To appear.

JEWELL, N. P., MALANI, H. M. and VITTINGHOFF, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *J. Amer. Statist. Assoc.* **89** 7–18.

POLLARD, D. (1989). Asymptotics via empirical processes (with discussion). *Statist. Sci.* **4** 341–366.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. IMS, Hayward, CA.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

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

SHIBOSKI, S. C. and JEWELL, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *J. Amer. Statist. Assoc.* **87** 360–372.

SILVERMAN, B. M. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.

VAN DER VAART, A. W. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204.

VAN DER VAART, A. W. and WELLNER, J. A. (1992). Existence and consistency of maximum likelihood in upgraded mixture models. *J. Multivariate Anal.* **43** 133–146.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

WONG, W. H. and SEVERINI, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *Ann. Statist.* **19** 603–632.

DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242