PENALIZED LIKELIHOOD HAZARD ESTIMATION

by

Chong Gu
Purdue University

Technical Report #91-58

Department of Statistics
Purdue University

October 1991

◇

Penalized Likelihood Hazard Estimation

CHONG GU*

Purdue University

Abstract

An asymptotic analysis of penalized likelihood hazard estimation using censored life time data is presented. The counting process interpretation of censored data and the associated martingale structure are employed. Asymptotic convergence rates in a certain symmetrized Kullback-Leibler and in a related mean square error are obtained. A computable adaptive estimator is proposed and is shown to share the same asymptotic convergence rates as the original estimator. An example is also provided.

AMS 1991 subject classifications. Primary 62G07; secondary 62G20, 65D07, 65D10, 41A15, 41A25.

Key words and phrases. Censored data, hazard, Kullback-Leibler, rate of convergence, smoothing.

1 Introduction

Censored life time data are common in life testing, medical follow up and other studies. Let T_i be the life time of an item and C_i be the censoring time beyond which the item is dropped from the study. One observes (X_i, δ_i), $i = 1, \cdots, n$, where $X_i = \min(T_i, C_i)$ and $\delta_i = I_{[T_i \le C_i]}$. Assume that T_i follow a common survival function $S(t) = \text{Prob}(T > t)$. Of interest is the estimation of the hazard function $\lambda(t) = -d \log S(t)/dt$.

Conventional estimators of $\lambda(t)$ include various parametric maximum likelihood estimators and the constraint-free nonparametric maximum likelihood delta sum corresponding to the Kaplan-Meier estimator of the survival function; see, e.g., Kalbfleisch and Prentice (1980). Parametric

*Research supported by NSF under Grant DMS-9101730.

estimators are restrictive, while the delta sum is "unreal". In between the two extremes, estimators with nonrestrictive constraints such as the penalized likelihood estimators provide a proper balance between regularity and adaptiveness in the estimation. As a general method, penalized likelihood method estimates a function of interest η via the minimizer of $L(\eta|\text{data}) + \lambda J(\eta)$, where the L, usually a minus log likelihood, measures the lack of fit of η to the data, the J, usually a quadratic functional, measures the roughness or irregularity of η, and the smoothing parameter λ, a positive constant not to be confused with the hazard function, controls the tradeoff between the smoothness and the goodness-of-fit of the estimator. Penalized likelihood method was introduced by Good and Gaskins (1971) in the context of nonparametric probability density estimation. Its use in hazard estimation was proposed by Anderson and Senthilselvan (1980), Bartoszynski, Brown, McBride and Thompson (1981), and O'Sullivan (1988). Cox and O'Sullivan (1990) conducted a general asymptotic analysis of penalized likelihood estimators, of which O'Sullivan's (1988) hazard estimator is a special case.

The purpose of this article is to conduct an asymptotic analysis of O'Sullivan's (1988) hazard estimator in a manner different from that of Cox and O'Sullivan (1990). The Cox-O'Sullivan analysis treats density estimation, regression, and hazard estimation all as general function estimation problems, yielding convergence results in terms of functional space norms. In contrast, the current analysis and its parallels (Gu and Qiu 1991a, b) adapt to the specific stochastic structures in different problems, leading to problem-specific convergence results under problem-specific conditions. The general theoretical framework, however, is common in all these parallel analyses. The specific stochastic structure in the current analysis features the counting process interpretation of censored life time data and the associated martingale structure; see, e.g., Fleming and Harrington (1991, Chapters 1-2).

The development in this article is organized as follows. Section 2 defines O'Sullivan's (1988) estimator and conducts preliminary analysis: In §2.1, the estimator to be analyzed is formally formulated and its existence is discussed. In §2.2, a symmetrized Kullback-Leibler is derived under the counting process framework to assess the estimation precision, and the martingale structure of the data is reviewed for later reference. In §2.3, the smoothness conditions characterizing the roughness penalty are discussed. Section 3 calculates the asymptotic convergence rates of the estimator in the symmetrized Kullback-Leibler and the related mean square error: In §3.1, a linear

approximation is analyzed. In §3.2, the distance between the estimator and the linear approximation is calculated, and the convergence rates of the estimator are obtained. In §3.3, a numerically computable semiparametric adaptive estimator is proposed and analyzed. Section 4 presents an example.

2 Formulation and preliminaries

2.1 Penalized likelihood estimation

Consider independent observations (X_i, δ_i), $i = 1, \cdots, n$, and assume independent censorship. Assume $\lambda(t) > 0$ wherever $\tilde{S}(t) = \text{Prob}(X_i \geq t) > 0$ and let $\eta(t) = \log \lambda(t)$. In the remaining of the article, I shall only use e^η to indicate the hazard and reserve the symbol λ exclusively for the smoothing parameter. Let $f(t) = e^{\eta(t)} S(t)$ be the probability density of T_i. The likelihood of the data is

$$\prod_{i=1}^{n} \{S(X_i)^{1-\delta_i} f(X_i)^{\delta_i}\} = \prod_{i=1}^{n} \{S(X_i) e^{\delta_i \eta(X_i)}\}.$$

Note that $S(t) = \exp(-\int_0^t e^\eta)$. The penalized likelihood estimate of η is defined as the minimizer of the functional

$$-\frac{1}{n} \sum_{i=1}^{n} \{\delta_i \eta(X_i) - \int_0^{X_i} e^\eta\} + \frac{\lambda}{2} J(\eta) \tag{2.1}$$

in a Hilbert space $\mathcal{H}$, where the first term is the minus log likelihood. The J is taken as a square norm in $\mathcal{H}$ or a square seminorm with a finite dimensional null space $J_\perp \subset \mathcal{H}$, where a finite dimensional $J_\perp$ prevents interpolation, the conceptual equivalent of a delta sum. Evaluation $[t]\eta = \eta(t)$ is assumed to be continuous in $\eta \in \mathcal{H}$, $\forall t \in \{t : \tilde{S} > 0\}$, which is necessary for (2.1) to be continuous in its argument η. This formulation is slightly more general than that of O'Sullivan (1988). An example is given in Section 4.

Assume that $\eta(t)$ is continuous in t, $\forall \eta \in \mathcal{H}$. By the Riemann sum approximations of $\int_0^{X_i} e^\eta$ and the continuity of evaluation, (2.1) is continuous in η. Now

$$\int e^{\alpha\eta_1 + \beta\eta_2} \leq \{\int e^{\eta_1}\}^\alpha \{\int e^{\eta_2}\}^\beta = \exp\{\alpha \log \int e^{\eta_1} + \beta \log \int e^{\eta_2}\} \leq \alpha \int e^{\eta_1} + \beta \int e^{\eta_2}$$

for $\alpha, \beta \in (0, 1)$, $\alpha + \beta = 1$, where the first (Holder's) inequality is strict unless $e^{\eta_1} \propto e^{\eta_2}$ and the second is strict unless $\int e^{\eta_1} = \int e^{\eta_2}$, so (2.1) is strictly convex in η.

Theorem 2.1 *The minimizer of (2.1) exists in $\mathcal{H}$ and is unique whenever it exists in $J_\perp$.*

The theorem is simply a corollary of Theorem 3.1 of Gu and Qiu (1991a).

2.2 Martingale structure

Let $N(t) = I_{[X \le t, \delta=1]}$. Under independent censorship, the quantity $e^{\eta(t)}dt$ is the conditional probability that $N(t)$ make a jump in $[t, t+dt)$ given that $X \ge t$. Letting η_0 be the true hazard and $\hat{\eta}$ the estimate, it is easy to show that the symmetrized Kullback-Leibler between two Bernoulli distributions with failure probabilities $e^{\eta_0(t)}dt$ and $e^{\hat{\eta}(t)}dt$ is $(e^{\hat{\eta}(t)} - e^{\eta_0(t)})(\hat{\eta}(t) - \eta_0(t))dt + O((dt)^2)$. Weighting by the probability $\tilde{S}(t) = \text{Prob}(X \ge t)$ and accumulating over $[0, \infty)$,

$$\text{SKL}(\eta_0, \hat{\eta}) = \int_0^\infty (e^{\hat{\eta}} - e^{\eta_0})(\hat{\eta} - \eta_0)\tilde{S} \tag{2.2}$$

defines an appropriate measure for assessing the estimation precision. Note that SKL is not a distance in the usual sense. Nevertheless, a quadratic norm $V(\eta) = \int_0^\infty \eta^2 e^{\eta_0}\tilde{S}$ defines a distance $V(\hat{\eta} - \eta_0)$ which approximates $\text{SKL}(\eta_0, \hat{\eta})$. Note that $e^{\eta_0(t)}\tilde{S}(t)dt$ is the probability that an item fails in $[t, t+dt)$, so $V(\hat{\eta} - \eta_0)$ is actually a properly weighted mean square error.

Let $Y(t) = I_{[X \ge t]}$ and $A(t) = \int_0^t Y(u)e^{\eta_0(u)}du$. Under independent censorship, $M(t) = N(t) - A(t)$ is a martingale. $EM(t) = 0$ and $EM^2(t) = EA(t) = \int_0^t e^{\eta_0}\tilde{S}$. Given any deterministic continuous function h on $[0, \infty)$ (so it is locally bounded predictable), the Stieltjes integral $\int_0^t h(u)dM(u)$ is also a martingale so long as $\int_0^\infty h^2 e^{\eta_0}\tilde{S} < \infty$. It follows that

$$E\int_0^t h\,dN - \int_0^t h\,e^{\eta_0}\tilde{S} = E\int_0^t h dM = 0 \tag{2.3}$$

and

$$E\{\int_0^t h\,dM\}^2 = E\int_0^t h^2 dA = \int_0^t h^2 e^{\eta_0}\tilde{S}. \tag{2.4}$$

Further,

$$\begin{aligned} E\{\int_0^t h\,dN - \int_0^t h\,e^{\eta_0}\tilde{S}\}^2 &= E\{\int_0^t h\,d(N-A) + \int_0^t h\,e^{\eta_0}(Y - \tilde{S})\}^2 \\ &= E\{\int_0^t h\,dM\}^2 + E\{\int_0^t h\,e^{\eta_0}(Y - \tilde{S})\}^2, \end{aligned} \tag{2.5}$$

where $E\int_0^t h\,dM \int_0^t h\,e^{\eta_0}(Y - \tilde{S}) = 0$ since $\int_0^t h\,e^{\eta_0}(Y - \tilde{S})$ is predictable. Note that $\delta\eta(X) = \int_0^\infty \eta dN$ and $\int_0^X e^\eta = \int_0^\infty Y e^\eta$. The functional (2.1) shall be written as

$$-\frac{1}{n}\sum_{i=1}^n \{\int \eta dN_i - \int Y_i e^\eta\} + \frac{\lambda}{2}J(\eta) \tag{2.6}$$

for later reference. From now on, integrals with unspecified limits shall be understood as over $[0, \infty)$.

The results quoted in this subsection are mainly taken from Fleming and Harrington (1991, §2.7). See also Gill (1984).

2.3 Smoothness assumptions

Assume that $e^{\eta_0(t)}\tilde{S}(t)$ decays fast enough as $t \to \infty$ so that $V(\eta) = \int \eta^2 e^{\eta_0}\tilde{S} < \infty$ for $\eta \in \mathcal{H}$. $V(\eta)$ defines a statistically interpretable metric in $\mathcal{H}$ as discussed in §2.2. The nonrestrictive constraints imposed by $\lambda J(\eta)$, or the smoothness of functions in $\mathcal{H}$, shall be characterized via an eigenvalue analysis of J with respect to V.

A bilinear form B is said to be completely continuous with respect to another bilinear form A, if for any $\epsilon > 0$, there exist finite number of linear functionals $l_1, \cdots, l_k$ such that $l_j(\eta) = 0$, $j = 1, \cdots, k$, implies that $B(\eta) \leq \epsilon A(\eta)$; see Weinberger (1974, §3.3). To possibly achieve noise reduction in estimation, the effective model space dimension has to be kept finite, while to make the estimation nonrestrictive, the effective model space dimension should be expandable as more data become available. Penalized likelihood method just tries to implement this, where for fixed λ the dimension may be kept down via keeping λJ bounded and the dimension expansion may be achieved by letting $\lambda \to 0$ as $n \to \infty$. To make this possible the following assumption must be made.

Assumption A.1. V is completely continuous with respect to J.

A.1 is equivalent to assuming that V is completely continuous with respect to $(V + J)$. Under A.1, using Theorem 3.1 of Weinberger (1974, p.52), it can be shown that there exist $\phi_\nu \in \mathcal{H}$ and $0 \leq \rho_\nu \uparrow \infty$, $\nu = 1, 2, \cdots$, such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu,\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta; see Gu and Qiu (1991, §4). The notion of smoothness is characterized by the rate of growth of ρ_ν.

Assumption A.2. $\rho_\nu = c_\nu \nu^r$, where $r > 1$, $c_\nu \in (\beta_1, \beta_2)$, and $0 < \beta_1 < \beta_2 < \infty$.

The asymptotic convergence rates of the estimators directly depend on r.

3 Asymptotic analysis

3.1 Linear approximation

Assume $\eta_0 \in \mathcal{H}$. Let η_1 be the minimizer of the quadratic functional

$$-\frac{1}{n}\sum_{i=1}^{n}\{\int \eta dN_i - \int \eta Y_i e^{\eta_0}\} + \frac{1}{2}V(\eta-\eta_0) + \frac{\lambda}{2}J(\eta). \tag{3.1}$$

Write $\eta = \sum_\nu \eta_\nu \phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$, where $\eta_\nu = V(\eta, \phi_\nu)$ are the Fourier coefficients of η with basis ϕ_ν. Substituting these into (3.1) and solving for $\eta_{\nu,1}$, one obtains $\eta_{\nu,1} = (\beta_\nu + \eta_{\nu,0})/(1+\lambda\rho_\nu)$, where $\beta_\nu = (1/n)\sum_{i=1}^n \int \phi_\nu dM_i$. By (2.3), (2.4), and that $\int \phi_\nu^2 e^{\eta_0}\tilde{S} = V(\phi_\nu) = 1$, $E\beta_\nu = 0$ and $E\beta_\nu^2 = n^{-1}$. It then follows that

$$\begin{aligned} EV(\eta_1-\eta_0) &= E\sum_{i=1}^{n}(\eta_{\nu,1}-\eta_{\nu,0})^2 = O(n^{-1}\lambda^{-1/r}+\lambda) \\ E\lambda J(\eta_1-\eta_0) &= E\lambda\sum_{i=1}^{n}\rho_\nu(\eta_{\nu,1}-\eta_{\nu,0})^2 = O(n^{-1}\lambda^{-1/r}+\lambda), \end{aligned}$$

as $n \to \infty$ and $\lambda \to 0$; see Gu and Qiu (1991, Theorem 4.1). See also Silverman (1982, §6).

Theorem 3.1 *Under A.1 and A.2, as $n \to \infty$ and $\lambda \to 0$, $V(\eta_1-\eta_0) = O_p(n^{-1}\lambda^{-1/r}+\lambda)$ and $\lambda J(\eta_1-\eta_0) = O_p(n^{-1}\lambda^{-1/r}+\lambda)$.*

3.2 Approximation error and main result

Let $L(\eta) = -(1/n)\sum_{i=1}^n\{\int \eta dN_i - \int Y_i e^\eta\}$ and $B_{\eta,h}(\alpha) = L(\eta+\alpha h) + (\lambda/2)J(\eta+\alpha h)$. It can be shown that

$$0 = \dot{B}_{\hat{\eta},\hat{\eta}-\eta_1}(0) = -\frac{1}{n}\sum_{i=1}^{n}\{\int(\hat{\eta}-\eta_1)dN_i - \int(\hat{\eta}-\eta_1)Y_i e^{\hat{\eta}}\} + \lambda J(\hat{\eta},\hat{\eta}-\eta_1). \tag{3.2}$$

Similarly, define $L_1(\eta) = (1/n)\sum_{i=1}^n\{\int \eta dN_i - \int \eta Y_i e^{\eta_0}\} + (1/2)V(\eta-\eta_0)$ and $C_{\eta,h}(\alpha) = L_1(\eta+\alpha h) + (\lambda/2)J(\eta+\alpha h)$. It follows that

$$0 = \dot{C}_{\eta_1,\hat{\eta}-\eta_1}(0) = -\frac{1}{n}\sum_{i=1}^{n}\{\int(\hat{\eta}-\eta_1)dN_i - \int(\hat{\eta}-\eta_1)Y_i e^{\eta_0}\} + V(\eta_1-\eta_0,\hat{\eta}-\eta_1) + \lambda J(\eta_1,\hat{\eta}-\eta_1). \tag{3.3}$$

Equating (3.2) and (3.3), some algebra yields

$$\int(\hat{\eta}-\eta_1)(e^{\hat{\eta}}-e^{\eta_1})\bar{Y} + \lambda J(\hat{\eta}-\eta_1) = V(\eta_1-\eta_0,\hat{\eta}-\eta_1) - \int(\hat{\eta}-\eta_1)(e^{\eta_1}-e^{\eta_0})\bar{Y}, \tag{3.4}$$

where $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$. One needs the following assumptions to proceed.

Assumption A.3. For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and η_1, $\exists c_1, c_2 \in (0, \infty)$ such that $c_1 e^{\eta_0(t)} \leq e^{\eta(t)} \leq c_2 e^{\eta_0(t)}$ uniformly on $\{t : \tilde{S}(t) > 0\}$.

A.3 assures the equivalence of the V distance and the SKL in B_0.

Assumption A.4. $\exists c_3 < \infty$ such that $\int \phi_\nu^2 e^{\eta_0} \tilde{S}^{1/2} \leq c_3$, $\forall \nu$.

A.4 requires a faster decay of $\tilde{S}$ than merely appropriate for defining V. By A.3,

$$c_1 \int (\hat{\eta} - \eta_1)^2 e^{\eta_0} \bar{Y} \leq \int (\hat{\eta} - \eta_1)(e^{\hat{\eta}} - e^{\eta_1}) \bar{Y}. \tag{3.5}$$

Writing $\hat{\eta} = \sum_\nu \hat{\eta}_\nu \phi_\nu$ and $\eta_1 = \sum_\nu \eta_{\nu,1} \phi_\nu$,

$$\begin{aligned}
& |\int (\hat{\eta} - \eta_1)^2 e^{\eta_0} \bar{Y} - V(\hat{\eta} - \eta_1)| \\
= \; & |\sum_\nu \sum_\mu (\hat{\eta}_\nu - \eta_{\nu,1})(\hat{\eta}_\mu - \eta_{\mu,1}) \int \phi_\nu \phi_\mu e^{\eta_0} (\bar{Y} - \tilde{S})| \\
\leq \; & \{\sum_\nu \sum_\mu (1 + \lambda \rho_\nu)(1 + \lambda \rho_\mu)(\hat{\eta}_\nu - \eta_{\nu,1})^2 (\hat{\eta}_\mu - \eta_{\mu,1})^2\}^{1/2} \\
& \{\sum_\nu \sum_\mu (1 + \lambda \rho_\nu)^{-1} (1 + \lambda \rho_\mu)^{-1} \{\int \phi_\nu \phi_\mu e^{\eta_0} (\bar{Y} - \tilde{S})\}^2\}^{1/2} \\
= \; & (V + \lambda J)(\hat{\eta} - \eta_1) O_p(n^{-1/2} \lambda^{-1/r}),
\end{aligned} \tag{3.6}$$

where Cauchy-Schwartz, that

$$E\{\int \phi_\nu \phi_\mu e^{\eta_0} (\bar{Y} - \tilde{S})\}^2 \leq \int \phi_\nu^2 e^{\eta_0} \tilde{S}^{1/2} \int \phi_\mu^2 e^{\eta_0} \tilde{S}^{-1/2} E(\bar{Y} - \tilde{S})^2 \leq c_3^2 / n, \tag{3.7}$$

and that $\sum_\nu (1 + \lambda \rho_\nu)^{-1} = O(\lambda^{-1/r})$ (Gu and Qiu, 1991a, Lemma 4.2) are used. Similarly,

$$\int (\hat{\eta} - \eta_1)(e^{\eta_1} - e^{\eta_0}) \bar{Y} = c \int (\eta_1 - \eta_0)(\hat{\eta} - \eta_1) e^{\eta_0} \bar{Y} \tag{3.8}$$

where $c \in [c_1, c_2]$, and

$$|\int (\eta_1 - \eta_0)(\hat{\eta} - \eta_1) e^{\eta_0} \bar{Y} - V(\eta_1 - \eta_0, \hat{\eta} - \eta_1)| = (V + \lambda J)^{1/2} (\eta_1 - \eta_0)(V + \lambda J)^{1/2} (\hat{\eta} - \eta_1) O_p(n^{-1/2} \lambda^{-1/r}). \tag{3.9}$$

Combining (3.4) – (3.9) and letting $n\lambda^{2/r} \to \infty$,

$$(c_1 V + \lambda J)(\hat{\eta} - \eta_1)(1 + o_p(1)) \leq |c - 1| V(\eta_1 - \eta_0, \hat{\eta} - \eta_1) + (V + \lambda J)^{1/2} (\eta_1 - \eta_0)(V + \lambda J)^{1/2} (\hat{\eta} - \eta_1) o_p(1). \tag{3.10}$$

Theorem 3.2 *Under A.1 - A.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(\hat{\eta} - \eta_1) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ *and.* $\lambda J(\hat{\eta} - \eta_1) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

The proof of the theorem follows from (3.10), Cauchy-Schwartz, and Theorem 3.1.

Theorem 3.3 *Under A.1 - A.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, $\lambda J(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, *and* $\mathrm{SKL}(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

Theorem 3.3 is a direct consequence of Theorems 3.1, 3.2, and Assumption A.3.

3.3 Semiparametric adaptive estimator

The space $\mathcal{H}$ is infinite dimensional and the estimator $\hat{\eta}$ is in general not computable. To make the procedure practically applicable, appropriate finite dimensional approximation of $\mathcal{H}$ is needed. O'Sullivan (1988) calculated the minimizer of (2.1) in a function space spanned by the B-spline bases on a finite interval. In this subsection, I shall propose a data-adaptive semiparametric estimator and analyze its asymptotic convergence.

Given a square norm in $J_\perp$, $\mathcal{H}$ has a tensor sum decomposition such that J is a square norm in $\mathcal{H} \ominus J_\perp$. A Hilbert space in which evaluation is continuous is known as a reproducing kernel Hilbert space possessing a reproducing kernel, a positive-definite bivariate function R with the reproducing property that $\langle R(t,\cdot), \eta\rangle = \eta(t)$, where $\langle\cdot,\cdot\rangle$ is the inner product in the space; see, e.g., Wahba (1990, Chapter 1). Let R_J be the reproducing kernel in the space $\mathcal{H} \ominus J_\perp$ with J as the inner product. The proposed adaptive estimator is the minimizer $\hat{\eta}_n$ of (2.1) in $\mathcal{H}_n = J_\perp \oplus \{R_J(X_i,\cdot), i = 1,\cdots,n\}$. Theorem 2.1 remains valid when $\mathcal{H}$ is replaced by $\mathcal{H}_n$.

Let $h \in \mathcal{H} \ominus \mathcal{H}_n \subset \mathcal{H} \ominus J_\perp$. It follows that $h(X_i) = J(R_J(X_i,\cdot), h) = 0$. So $\sum_i^n \int h^2 dN_i = \sum_{i=1}^n \delta_i h^2(X_i) = 0$. By (2.3) - (2.5), $E\{\int \phi_\nu\phi_\mu d\bar{N} - \int \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}\} = 0$ and

$$E\{\int \phi_\nu\phi_\mu d\bar{N} - \int \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}\}^2 = \frac{1}{n}\int \phi_\nu^2\phi_\mu^2 e^{\eta_0}\tilde{S} + E\{\int \phi_\nu\phi_\mu e^{\eta_0}(\bar{Y} - \tilde{S})\}^2, \qquad (3.11)$$

where $\bar{N} = (1/n)\sum_{i=1}^n N_i$.

Assumption A.5. $\exists c_4 < \infty$ such that $\int \phi_\nu^2\phi_\mu^2 e^{\eta_0}\tilde{S} \le c_4, \forall \nu, \mu$.

Lemma 3.1 *Under A.1, A.2, A.4 and A.5, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(h) = \lambda J(h) o_p(1)$ *for* $h \in \mathcal{H} \ominus \mathcal{H}_n$.

Proof: Similar to (3.6),

$$V(h) = |\sum_{\nu}\sum_{\mu} h_\nu h_\mu \{\int \phi_\nu \phi_\mu d\bar{N} - \int \phi_\nu \phi_\mu e^{\eta_0}\tilde{S}\}| = (V+\lambda J)(h) O_p(n^{-1/2}\lambda^{-1/r}),$$

where (3.11), (3.7) and A.5 are used to bound $E\{\int \phi_\nu\phi_\mu d\bar{N} - \int \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}\}^2$. □

Let η_n be the projection of $\hat{\eta}$ onto $\mathcal{H}_n$. Note that $\dot{B}_{\hat{\eta},\hat{\eta}-\eta_n}(0) = 0$ and that $J(\eta_n, \hat{\eta}-\eta_n) = 0$. It follows that

$$\lambda J(\hat{\eta}-\eta_n) = \int (\hat{\eta}-\eta_n) d\bar{M} - \int (\hat{\eta}-\eta_n)(e^{\hat{\eta}} - e^{\eta_0})\bar{Y}, \tag{3.12}$$

where $\bar{M} = (1/n)\sum_{i=1}^n M_i$. Using the technique used in (3.6),

$$|\int (\hat{\eta}-\eta_n) d\bar{M}| = |\sum_{\nu}(\hat{\eta}_\nu - \eta_{\nu,n}) \int \phi_\nu d\bar{M}| = (V+\lambda J)^{1/2}(\hat{\eta}-\eta_n) O_p(n^{-1/2}\lambda^{-1/2r}). \tag{3.13}$$

Similar to (3.8) and (3.9), letting $n\lambda^{2/r} \to \infty$ and using A.3 and Lemma 3.1,

$$|\int (\hat{\eta}-\eta_n)(e^{\hat{\eta}} - e^{\eta_0})\bar{Y}| = (\lambda J)^{1/2}(\hat{\eta}-\eta_n)(V+\lambda J)^{1/2}(\hat{\eta}-\eta_0) o_p(1) \tag{3.14}$$

Theorem 3.4 *Under A.1 - A.5, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, $\lambda J(\hat{\eta}-\eta_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ and $V(\hat{\eta}-\eta_n) = o_p(n^{-1}\lambda^{1/r} + \lambda)$.*

The proof of Theorem 3.4 follows from (3.12) – (3.14) and Theorem 3.3.

I shall now calculate $V(\hat{\eta}_n - \eta_n)$. From $\dot{B}_{\hat{\eta}_n,\hat{\eta}_n-\eta_n}(0) = \dot{B}_{\hat{\eta},\hat{\eta}_n-\hat{\eta}}(0) = 0$, noting that $J(\hat{\eta} - \eta_n, \eta_n) = J(\hat{\eta}-\eta_n, \hat{\eta}_n) = 0$ so $J(\hat{\eta}, \hat{\eta}-\hat{\eta}_n) = J(\hat{\eta}-\eta_n) + J(\eta_n, \eta_n - \hat{\eta}_n)$, it can be shown that

$$\begin{aligned}\int (\hat{\eta}_n-\eta_n)(e^{\hat{\eta}_n} - e^{\eta_n})\bar{Y} + \lambda J(\hat{\eta}_n-\eta_n) + \lambda J(\hat{\eta}-\eta_n) \;&=\; \int (\hat{\eta}-\eta_n) d\bar{M} + \int (\hat{\eta}_n-\eta_n)(e^{\hat{\eta}} - e^{\eta_n})\bar{Y} \\ &\quad + \int (\hat{\eta}-\eta_n)(e^{\eta_0} - e^{\hat{\eta}})\bar{Y}\end{aligned} \tag{3.15}$$

Modify A.3 to include η_n and $\hat{\eta}_n$ in B_0. It follows that, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,

$$c_1 V(\hat{\eta}_n-\eta_n) + (V+\lambda J)(\hat{\eta}_n-\eta_n) o_p(1) \le \int (\hat{\eta}_n-\eta_n)(e^{\hat{\eta}_n} - e^{\eta_n})\bar{Y}, \tag{3.16}$$

$$|\int (\hat{\eta}_n-\eta_n)(e^{\hat{\eta}} - e^{\eta_n})\bar{Y}| = (V+\lambda J)^{1/2}(\hat{\eta}_n-\eta_n)(\lambda J)^{1/2}(\hat{\eta}-\eta_n) o_p(1), \tag{3.17}$$

and

$$|\int (\hat{\eta}-\eta_n)(e^{\eta_0} - e^{\hat{\eta}})\bar{Y}| = (V+\lambda J)^{1/2}(\hat{\eta}-\eta_0)(\lambda J)^{1/2}(\hat{\eta}-\eta_n) o_p(1). \tag{3.18}$$

Combining (3.15) – (3.18) and (3.13), and substituting in the results of Theorems 3.3 and 3.4,

$$(c_1 V + \lambda J)(\hat{\eta}_n - \eta_n)(1 + o_p(1)) \leq (V + \lambda J)^{1/2}(\hat{\eta}_n - \eta_n) o_p(n^{-1/2}\lambda^{-1/2r} + \lambda^{1/2}) + O_p(n^{-1}\lambda^{-1/r} + \lambda). \tag{3.19}$$

This proves the following theorem.

Theorem 3.5 *Under A.1 - A.5, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $\lambda J(\hat{\eta}_n - \eta_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ *and* $V(\hat{\eta}_n - \eta_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

The next theorem follows from Theorems 3.3, 3.4, 3.5, and Assumption A.3.

Theorem 3.6 *Under A.1 - A.5, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(\hat{\eta}_n - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, $\lambda J(\hat{\eta}_n - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, *and* $\mathrm{SKL}(\hat{\eta}_n - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

The results of this subsection remain valid when $\mathcal{H}_n$ is reduced to $J_\perp \oplus \{R_J(X_i, \cdot), \delta_i = 1\}$.

4 Example

Consider the following example. Suppose $\exists\, T_c < \infty$ such that $\tilde{S}(T_c) > 0$ and $\tilde{S}(T_c + 0) = 0$, which amounts to having a type I censoring scheme active at time T_c, though not necessarily exclusively. Without loss of generality let $T_c = 1$. Take $J(\eta) = \int_0^1 \ddot{\eta}^2$ and $\mathcal{H} = \{\eta : J(\eta) < \infty\}$. Assuming that η_0 is bounded from above and below on $[0, 1]$, it can be shown that A.1 and A.2 are satisfied with $r = 4$ in A.2; see, e.g., Silverman (1982, p.802). A.3 is asking that the members of B_0 stay in a strip centered at η_0 with a fixed width, which appears reasonable though not directly verifiable. A.4 is trivial in this example since $V(\phi_\nu) = 1$ and $\tilde{S}^{-1/2}$ has an upper bound $\tilde{S}(1)^{-1/2} < \infty$. In general, A.5 can not be directly verified because ϕ_ν are not available in explicit forms. In a suggestive special case where $e^{\eta_0}\tilde{S} = 1$ on $[0, 1]$ and $\mathcal{H}$ is reduced to the periodic restriction of $\{\eta : J(\eta) < \infty\}$, however, ϕ_ν are sines and cosines and hence are uniformly bounded, and in turn A.5 follows. To construct $\mathcal{H}_n$ one needs only to identify $J_\perp$ and R_J. $J_\perp = \{1, t\}$. If the square norm in $J_\perp$ is taken as $\eta^2(0) + \dot{\eta}^2(0)$, then $\mathcal{H} \ominus J_\perp = \{\eta : \eta(0) = \dot{\eta}(0) = 0, J(\eta) < \infty\}$, and $R_J(t, s) = \int_0^1 (t - u)_+(s - u)_+ du$, where $(\cdot)_+ = \max(\cdot, 0)$; see, e.g., Gu and Qiu (1991a, §2). The numerical calculation of $\hat{\eta}_n$ shall be studied elsewhere.

References

Anderson, J. A. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. *J. Roy. Statist. Soc.* B **42**, 322 – 327.

Bartoszynski, R., Brown, B. W., McBride, C. M., and Thompson, J. R. (1981). Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. *Ann. Statist.* **9**, 1050 – 1060.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18**, 1676 – 1695.

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

Gill, R. D. (1984). Understanding Cox's regression model: a martingale approach. *J. Amer. Statist. Assoc.* **79**, 441 – 447.

Gu, C. and Qiu, C. (1991a). Smoothing spline density estimation: Theory. Technical Report 91-19, Purdue University, Dept. of Statistics.

—— (1991b). Asymptotic analysis of penalized likelihood regression. Technical Report 91-50, Purdue University, Dept. of Statistics.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9**, 363 – 379.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795 – 810.

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 59. SIAM, Philadelphia.

Weinberger, H. F. (1974). *Variational Methods for Eigenvalue Approximation.* CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 15. SIAM, Philadelphia.