Statistica Sinica **6**(1996), 861-876

PENALIZED LIKELIHOOD HAZARD ESTIMATION: A GENERAL PROCEDURE

Chong Gu

Purdue University

Abstract: A general penalized likelihood hazard estimation procedure is formulated and an asymptotic theory developed. The life time data may be left-truncated, partly right-censored, and may come with a covariate. In the presence of a covariate, the modular model construction via tensor-product splines provides rich collections of hazard models, of which the proportional hazard model and a model of Zucker and Karr (1990) are special cases. The counting process interpretation of life time data and the associated martingale structure are employed in the analysis. Asymptotic convergence rates in a certain symmetrized Kullback-Leibler divergence and in a related mean square error are obtained. A computable adaptive estimate is proposed and is shown to share the same asymptotic convergence rates. A few examples are presented in some detail.

Key words and phrases: Covariate, hazard, Kullback-Leibler divergence, left-truncation, convergence rate, right-censoring, 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, Z_i be the (left) truncation time at which the item enters the study, and C_i be the (right) censoring time beyond which the item is dropped from the study, independent of each other. One observes $(Z_i, X_i, \delta_i, U_i)$, $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$, $\delta_i = I_{[T_i \leq C_i]}$, $Z_i < X_i$, and U_i is a covariate. Assume that $T_i|U_i$ follow a survival function $S(t,u) = \mathrm{Prob}(T > t|U = u)$. Of interest is the estimation of the hazard function $\lambda(t,u) = -\partial \log S(t,u)/\partial t$.

When the covariate U is absent, conventional estimates of $\lambda(t)$ include various parametric maximum likelihood estimates and the constraint-free nonparametric maximum likelihood delta sum corresponding to the Kaplan-Meier estimate of the survival function (see, e.g., Kalbfleisch and Prentice (1980)). Parametric estimates are restrictive, while the delta sum is “unreal”. In between the two extremes, estimates with nonrestrictive constraints such as the penalized likelihood estimates provide a proper balance between regularity and adaptiveness in the estimation. As a general method, the penalized likelihood method estimates a function of interest η via the minimizer of $L(\eta|\text{data}) + \lambda J(\eta)$, where

L, usually a minus log likelihood, measures the lack of fit of η to the data, 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 estimate. The 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), O'Sullivan (1988a), and Antoniadis (1989). Cox and O'Sullivan (1990) conducted a general asymptotic analysis of penalized likelihood estimates, of which O'Sullivan's (1988a) hazard estimate is a special case.

When the covariate U is present, a popular model is Cox's (1972) proportional hazard model, which assumes that $\lambda(t,u)=\lambda_0(t)\lambda_1(u)$. Cox's (1972) partial likelihood method treats $\lambda_0(t)$ as a nuisance, free of constraint, and imposes parametric models for $\lambda_1(u)$. O'Sullivan (1988b) substitutes splines for $\lambda_1(u)$ via penalized partial likelihood. Zucker and Karr (1990) considered a model of the form $\lambda(t,u)=\lambda_0(t)\lambda_1(\beta(t),u)$, where $\lambda_1(\beta(t),u)$ is parametric in u with a time-varying parameter $\beta(t)$, and estimated $\beta(t)$ via penalized partial likelihood.

In this article, smooth function models for $\lambda(t,u)$ on the product domain $\mathcal{T}\times\mathcal{U}$ of time and covariate are proposed via penalized full likelihood. By introducing a function decomposition of $\log\lambda(t,u)$, the models are made more general than but reducible to proportional hazard models and the model of Zucker and Karr (1990). When U is absent, the estimate reduces to that of O'Sullivan (1988a). The asymptotic convergence of the estimate and that of a computable adaptive estimate are studied via the approach of Gu and Qiu (1993) under the counting process interpretation of censored life time data and the associated martingale structure (cf. Fleming and Harrington (1991), Chapters 1-2). The computation of the adaptive estimate in the absence of U was studied in Gu (1994). An automatic algorithm for computing the adaptive estimate in the general setup, with data examples, was explored in Gu (1995).

The remainder of the article is organized as follows. Section 2 defines the estimate and conducts preliminary analysis: In §2.1, the estimate to be analyzed is formally formulated and its existence is discussed; in §2.2, a symmetrized Kullback-Leibler divergence 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 presents a few examples. Section 4 calculates the asymptotic convergence rates of the estimates in the symmetrized Kullback-Leibler and the related mean square error: In §4.1, a linear approximation of the estimate is analyzed; in §4.2, the distance between the estimate and the linear approximation is calculated, and the convergence rates of the estimate are obtained; in §4.3, a numerically computable semiparametric adaptive estimate is proposed and analyzed. Section 5 adds a few more details to the examples of Section 3 in the light of the theory. Section 6 contains a few remarks.

2. Formulation and Preliminaries

2.1. Penalized likelihood estimation

Consider independent observations $(Z_i, X_i, \delta_i, U_i)$, $i = 1, \ldots, n$, and assume independent censorship. Assume that U has a density $m(u) > 0$ on $\mathcal{U}$ and that $\lambda(t,u) > 0$ whenever $\tilde{S}(t,u) = \text{Prob}(Z < t \le X|U = u) > 0$, and let $\eta(t,u) = \log\lambda(t,u)$. In the remainder of the article, I only use e^η to indicate the hazard and reserve the symbol λ exclusively for the smoothing parameter. Let $f(t,u) = e^{\eta(t,u)}S(t,u)$ be the probability density of $T|U = u$. The likelihood of the data is

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

Note that $S(t,u) = \exp(-\int_0^t e^{\eta(s,u)}ds)$. A penalized likelihood estimate of η is defined as a minimizer of the functional

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\delta_i\eta(X_i,U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)}dt\Big\} + \frac{\lambda}{2}J(\eta) \qquad (2.1)$$

in a Hilbert space $\mathcal{H}$ on $\mathcal{T}\times\mathcal{U}$, where the first term is the minus log likelihood. 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 $[t,u]\eta = \eta(t,u)$ is assumed to be continuous in $\eta \in \mathcal{H}$, $\forall(t,u) \in \{(t,u) : \tilde{S}(t,u) > 0\}$, which is necessary for (2.1) to be continuous in its argument η. When $\mathcal{U}$ is a singleton, the formulation reduces to a slightly more general version of that of O'Sullivan (1988a). Examples will be given in Section 3.

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

$$\begin{aligned}\int e^{\alpha\eta_1+\beta\eta_2}dt &\le \Big\{\int e^{\eta_1}dt\Big\}^\alpha\Big\{\int e^{\eta_2}dt\Big\}^\beta = \exp\Big\{\alpha\log\int e^{\eta_1}dt + \beta\log\int e^{\eta_2}dt\Big\}\\ &\le \alpha\int e^{\eta_1}dt + \beta\int e^{\eta_2}dt\end{aligned}$$

for $\alpha, \beta \in (0,1)$, $\alpha + \beta = 1$, where the first (Holder's) inequality is strict unless $e^{\eta_1} \propto e^{\eta_2}$ on $(Z,X]\times\{U\}$ and the second is strict unless $\int e^{\eta_1}dt = \int e^{\eta_2}dt$ on $\{U\}$, so the likelihood part $L(\eta)$ of (2.1) is convex in η and the convexity is strict in any function space which keeps its dimension when restricted to $\cup_{i=1}^{n}\{(Z_i, X_i]\times\{U_i\}\}$.

Theorem 2.1. *A minimizer* $\hat{\eta}$ *of* (2.1) *exists in* $\mathcal{H}$ *whenever it uniquely exists in* $J_\perp$.

The theorem is simply a corollary of Theorem 4.1 of Gu and Qiu (1993). Note that although in the statement of Theorem 4.1 of Gu and Qiu (1993) $L(\eta)$ is assumed to be strictly convex, the proof essentially holds when the convexity is strict only in $J_\perp$.

The space $\mathcal{H}$ is infinite dimensional and the estimate $\hat{\eta}$ is in general not computable. To make the procedure practically applicable, appropriate finite dimensional approximation of $\mathcal{H}$ is needed. Below I propose a data-adaptive semiparametric estimate $\hat{\eta}_n$ in a certain finite dimensional space $\mathcal{H}_n \subset \mathcal{H}$, to be computed in practical applications. In Section 4, it will be shown that $\hat{\eta}$ and $\hat{\eta}_n$ share the same asymptotic convergence rates under appropriate conditions.

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 $\langle R((t,u),\cdot),\eta\rangle = \eta(t,u)$, where $\langle\cdot,\cdot\rangle$ is the inner product in the space (see, e.g., Aronszajn (1950) and 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. Take $\mathcal{H}_n = J_\perp \oplus \{R_J((X_i,U_i),\cdot),\delta_i = 1\}$; $\hat{\eta}_n$ is defined as a minimizer of (2.1) in $\mathcal{H}_n$. Theorem 2.1 remains valid when $\mathcal{H}$ is replaced by $\mathcal{H}_n$.

The specifications of R_J in the examples of Section 3 will be discussed in Section 5. The numerical calculation of $\hat{\eta}_n$ with automatic smoothing parameters can be found in a further work Gu (1995).

2.2. Martingale structure

Let $N(t) = I_{[X\le t,\delta=1]}$. Under independent censorship, the quantity $e^{\eta(t,u)}dt$ is the conditional probability that $N(t)$ makes a jump in $[t, t+dt)$ given that $t \le X$ and $U = u$. Letting η_0 be the true hazard and $\hat{\eta}$ the estimate, it is easy to show that the symmetrized Kullback-Leibler divergence between two Bernoulli distributions with failure probabilities $e^{\eta_0(t,u)}dt$ and $e^{\hat{\eta}(t,u)}dt$ is $(e^{\hat{\eta}(t,u)} - e^{\eta_0(t,u)})(\hat{\eta}(t,u) - \eta_0(t,u))dt + O((dt)^2)$. Weighting by the at-risk probability $\tilde{S}(t,u) = \text{Prob}(Z < t \le X|U = u)$ and accumulating over $\mathcal{T}$, $\int_{\mathcal{T}}(e^{\hat{\eta}}-e^{\eta_0})(\hat{\eta}-\eta_0)\tilde{S}$ defines a natural discrepancy measure conditional on $U = u$, and in turn

$$\text{SKL}(\eta_0,\hat{\eta}) = \int_{\mathcal{U}}\int_{\mathcal{T}}(e^{\hat{\eta}} - e^{\eta_0})(\hat{\eta} - \eta_0)\tilde{S}m \tag{2.2}$$

makes an appropriate measure for assessing the estimation precision, where $m(u) > 0$ is the density of U. Note that SKL is not a normed distance. Nevertheless, a quadratic norm $V(\eta) = \int_{\mathcal{U}}\int_{\mathcal{T}} \eta^2 e^{\eta_0}\tilde{S}m$ defines a distance $V(\hat{\eta} - \eta_0)$ which approximates $\mathrm{SKL}(\eta_0, \hat{\eta})$. Note that $e^{\eta_0(t,u)}\tilde{S}(t,u)dt$ is the probability that an item fails in $[t, t+dt)$ conditional on $U = u$, so $V(\hat{\eta} - \eta_0)$ is actually a properly weighted mean square error.

Let $Y(t) = I_{[Z<t\leq X]}$ be the at-risk process and $A(t) = \int_0^t Y(s)e^{\eta_0(s,U)}ds$. Under independent censorship, $M(t) = N(t) - A(t)$ is a martingale conditional on U and Z. $E[M(t)|U,Z] = 0$ and $E[M^2(t)|U,Z] = E[A(t)|U,Z] = \int_0^t e^{\eta_0}E[Y|U,Z]$. Given any deterministic function $h(t,u)$ continuous in t on $\mathcal{T}\times\mathcal{U}$ (so it is locally bounded predictable), the Stieltjes integral $\int_0^t h(s,U)dM(s)$ is also a martingale (conditional on U and Z) so long as $\int_{\mathcal{T}} h^2 e^{\eta_0}E[Y|U,Z] < \infty$, and in turn $E[\int_0^t hdM|U,Z] = 0$ and $E[\{\int_0^t hdM\}^2|U,Z] = \int_0^t h^2 e^{\eta_0}E[Y|U,Z]$. It follows that

$$E\int_0^t h\,dN - \int_{\mathcal{U}}\int_0^t h\,e^{\eta_0}\tilde{S}m = E\int_0^t hdM = 0 \tag{2.3}$$

and

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

Further,

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

where $E[\int_0^t h\,dM\{\int_0^t he^{\eta_0}Y - \int_{\mathcal{U}}\int_0^t he^{\eta_0}\tilde{S}m\}|U,Z] = 0$ because $\int_0^t he^{\eta_0}Y - \int_{\mathcal{U}}\int_0^t he^{\eta_0}\tilde{S}m$ is predictable. Note that $\delta\eta(X,U) = \int_{\mathcal{T}} \eta(t,U)dN(t)$ and $\int_Z^X e^{\eta} = \int_{\mathcal{T}} Ye^{\eta}$. The functional (2.1) shall be written as

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

for later reference, where $\eta_i(t) = \eta(t, U_i)$.

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 $V(\eta) = \int_{\mathcal{U}}\int_{\mathcal{T}} \eta^2 e^{\eta_0}\tilde{S}m < \infty$ 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 $l_1, \ldots, l_k$ such that $l_j(\eta) = 0$, $j = 1, \ldots, k$, 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. The 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 $0 \leq \rho_\nu \uparrow \infty$, $\nu = 1, 2, \ldots$, 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 (1993), §5). 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 $0 < \beta_1 < \beta_2 \leq \infty$.

The asymptotic convergence rates of the estimates directly depend on r. Technically, only a lower bound $\beta_1 > 0$ is required in the asymptotic analysis, but a nil upper bound $\beta_2 = \infty$ may imply nominal rates which are slower than the actual ones.

3. Examples

I present a few examples in this section and discuss assumptions A.1 and A.2. Assumptions A.3, A.4, and the specifications of R_J necessary for the calculation of $\hat{\eta}_n$ will be deferred to Section 5. These technically elementary examples are meant to illustrate the practical applicability of the method, complementing the rather abstract setting in which the theory is developed. Algorithms and data examples are to be found in Gu (1994, 1995).

Let $[T_0, T_1] = \cup_{u \in \mathcal{U}} \{\tilde{S}(t, u) > 0\}$, where T_0 and T_1 indicate the start and close time of the study. A type I censoring at some $t_c < \infty$ is almost always present in practical situations, explicitly or implicitly, so it appears reasonable to assume that $T_1 < \infty$. Without loss of generality, I shall take $T_0 = 0$, $T_1 = 1$, and $\mathcal{T} = [0, 1]$.

Example 1. *Singleton* $\mathcal{U}$. A singleton $\mathcal{U}$ characterizes the absence of a covariate. Take $J(\eta) = \int_0^1 \ddot{\eta}^2$ and $\mathcal{H} = \{\eta : J(\eta) < \infty\}$. It follows that $J_\perp = \{1, (\cdot)\}$. A minimizer of (2.1) exists whenever the maximum likelihood estimate of a parametric form $\eta(t) = \beta_0 + \beta_1 t$ uniquely exists. When $e^{\eta_0}\tilde{S}$ is bounded from zero and infinity on $[0,1]$ so $V(\eta) = \int_0^1 \eta^2 e^{\eta_0}\tilde{S}$ is equivalent to the L_2 norm, standard results (cf. Utreras (1981), Silverman (1982)) imply that A.1 and A.2 are satisfied, and $r = 4$ in A.2. This configuration was first considered by O'Sullivan (1988a). (See also Antoniadis (1989).)

I denote by $\{\psi_\nu\}$ the function sequence on $[0,1]$ satisfying $\int_0^1 \psi_\nu \psi_\mu = \delta_{\nu,\mu}$ and $\int_0^1 \ddot{\psi}_\nu \ddot{\psi}_\mu = \sigma_\nu \delta_{\nu,\mu}$, where $O(\nu^4) = \sigma_\nu \uparrow$. The first two entries are $\psi_1 = 1$ and $\psi_2 = 12^{1/2}(\cdot - .5)$ with $\sigma_1 = \sigma_2 = 0$.

Example 2. *Doubleton* $\mathcal{U}$. A doubleton $\mathcal{U}$, say $\mathcal{U} = \{1,2\}$, provides the simplest possible example of a categorical covariate. Functions on $\mathcal{U}$ are actually vectors in R^2. Define $\omega_1(u) = 2^{-1/2}$ and $\omega_2(u) = 2^{-1/2}(-1)^u$. ω_1 and ω_2 are orthonormal under the standard Euclidean norm. Let $\tilde{V}(\eta) = \sum_u \int_0^1 \eta^2$ and

$$\tilde{J}(\eta)=\sum_u \int_0^1 \ddot{\eta}^2=\int_0^1 (\ddot{\eta}(t,1)+\ddot{\eta}(t,2))^2/2+\int_0^1 (\ddot{\eta}(t,1)-\ddot{\eta}(t,2))^2/2=J_1(\eta)+J_2(\eta).$$

Set $\{\psi_\nu(t)\omega_j(u)\}$ as $\{\tilde{\phi}_\nu\}$ with $\tilde{\rho}_\nu = \tilde{J}(\tilde{\phi}_\nu)$ increasing. It is easy to verify that $\tilde{V}(\tilde{\phi}_\nu, \tilde{\phi}_\mu) = \delta_{\nu,\mu}$, $\tilde{J}(\tilde{\phi}_\nu, \tilde{\phi}_\mu) = \tilde{\rho}_\nu \delta_{\nu,\mu}$, and $\tilde{\rho}_\nu = O(\nu^4)$. As a matter of fact, $\tilde{J}_\perp = \{\psi_1\omega_1, \psi_1\omega_2, \psi_2\omega_1, \psi_2\omega_2\}$, J_1 defines a square norm in $\mathcal{H}_1 = \{\psi_\nu\omega_1\}_{\nu\geq 2}$, and J_2 defines a square norm in $\mathcal{H}_1 = \{\psi_\nu\omega_2\}_{\nu\geq 2}$. Now assume that $e^{\eta_0}\tilde{S}$ is bounded from zero and infinity on $[0,1] \times \{1,2\}$ and that $m(1), m(2) > 0$ so $V(\eta) = \int_{\mathcal{U}} \int_{\mathcal{T}} \eta^2 e^{\eta_0}\tilde{S}m$ is equivalent to $\tilde{V}(\eta)$. Also let $J(\eta) = \theta_1^{-1}J_1(\eta)+\theta_2^{-1}J_2(\eta)$, $\theta_1, \theta_2 > 0$, which is equivalent to $\tilde{J}(\eta)$. A.1 and A.2 are satisfied with $r = 4$ via standard arguments. A minimizer of (2.1) exists when the maximum likelihood estimate of the form $\eta(t,u) = \alpha_u + \beta_u t$ uniquely exists.

Note that J_1 penalizes the roughness of the average (over $\mathcal{U}$) log hazard, or the (time-axis) main effect, whereas J_2 penalizes the the roughness of the contrast log hazard, or the interaction. Setting $\theta_2 = 0+$ to effectively eliminate $\mathcal{H}_2$ and removing $\{\psi_2\omega_2\}$ from $\tilde{J}_\perp$, one obtains a proportional hazard model, which can be estimated by using $J = J_1$ in (2.1) and restricting $\mathcal{H}$ to $\{\psi_1\omega_1, \psi_1\omega_2, \psi_2\omega_1\} \oplus \mathcal{H}_1$, a subspace of $\tilde{J}_\perp \oplus \mathcal{H}_1 \oplus \mathcal{H}_2$ which characterizes functions of the form $\eta(t,u) = C + f_t + f_u$. When $\tilde{V}$ and V are equivalent, A.1 and A.2 (with $r = 4$) are satisfied in the proportional hazard model. A minimizer of (2.1) exists whenever the maximum likelihood estimate of the form $\eta(t,u) = \alpha_u + \beta t$ uniquely exists.

Example 3. $\mathcal{U} = [0,1]$. This describes a univariate continuous covariate. Let $\tilde{V}(\eta) = \int_0^1 \int_0^1 \eta^2$ and $\tilde{J}(\eta) = J_1(\eta) + J_2(\eta) + J_3(\eta) + J_4(\eta) + J_5(\eta)$, where $J_1 = \int_0^1 (\int_0^1 \ddot{\eta}_{tt} du)^2 dt$, $J_2 = \int_0^1 (\int_0^1 \ddot{\eta}_{uu} dt)^2 du$, $J_3 = \int_0^1 (\int_0^1 \eta^{(3)}_{ttu} du)^2 dt$, $J_4 = \int_0^1 (\int_0^1 \eta^{(3)}_{tuu} dt)^2 du$, and $J_5 = \int_0^1 \int_0^1 (\eta^{(4)}_{ttuu})^2 dtdu$. The sequence $\{\psi_\nu(t)\psi_\mu(u)\}$ is orthonormal under

$\tilde{V}$ and orthogonal under $\tilde{J}$. More precisely, J_1 defines a square norm in $\mathcal{H}_1 = \{\psi_\nu(t)\psi_1(u)\}_{\nu\geq 3}$, J_2 in $\mathcal{H}_2 = \{\psi_1(t)\psi_\nu(u)\}_{\nu\geq 3}$, J_3 in $\mathcal{H}_3 = \{\psi_\nu(t)\psi_2(u)\}_{\nu\geq 3}$, J_4 in $\mathcal{H}_4 = \{\psi_2(t)\psi_\nu(u)\}_{\nu\geq 3}$, J_5 in $\mathcal{H}_5 = \{\psi_\nu(t)\psi_\mu(u)\}_{\nu,\mu\geq 3}$, and the null space is $\tilde{J}_\perp = \{\psi_\nu(t)\psi_\mu(u)\}_{\nu,\mu=1,2}$. Taking $\{\sigma_\nu\sigma_\mu\}_{\nu,\mu\geq 3}$ in increasing order as $\{\tilde{\sigma}_\nu\}$ increases, it can be shown that $\tilde{\sigma}_\nu$ grows at a rate faster than $(\nu/\log\nu)^4$ but slower than ν^4 (cf. Wahba (1990), §12.1). Assuming that $e^{\eta_0}\tilde{S}m$ is bounded from above on $[0,1]\times[0,1]$ and taking $J(\eta) = \sum_{\beta=1}^{5}\theta_\beta^{-1}J_\beta(\eta)$ where $\theta_\beta > 0$, A.1 and A.2 follow standard arguments with a nil upper bound and $r = 4-\epsilon$, $\forall\epsilon > 0$.

Once again, a function decomposition is available for possible model simplifications. Two such simplifications are briefly described below, where $\mathcal{H}_5$ is eliminated so $r = 4$.

Taking $J(\eta)=\theta_1^{-1}J_1(\eta)+\theta_2^{-1}J_2(\eta)$ in $\mathcal{H}=\{\psi_1(t)\psi_1(u), \psi_1(t)\psi_2(u), \psi_2(t)\psi_1(u)\}\oplus\mathcal{H}_1\oplus\mathcal{H}_2$, one obtains a proportional hazard model. Such a model was considered by O'Sullivan (1988b), where the function component in $\{\psi_1(t)\psi_1(u), \psi_2(t)\psi_1(u)\}\oplus\mathcal{H}_1$ was treated as a nuisance and that in $\{\psi_1(t)\psi_2(u)\}\oplus\mathcal{H}_2$ estimated via penalized partial likelihood with J_2 as the penalty.

Taking $J(\eta) = \theta_1^{-1}J_1(\eta) + \theta_3^{-1}J_3(\eta)$ in $\mathcal{H} = \tilde{J}_\perp \oplus \mathcal{H}_1 \oplus \mathcal{H}_3$, one obtains a model of the form $\eta(t,u) = \alpha(t) + \beta(t)u$. Such a model was considered by Zucker and Karr (1990), where effectively the function component $\alpha(t) + .5\beta(t)$ in $\{\psi_1(t)\psi_1(u), \psi_2(t)\psi_1(u)\}\oplus\mathcal{H}_1$ was treated as a nuisance and the function component $\beta(t)(u-.5)$ in $\{\psi_1(t)\psi_2(u), \psi_2(t)\psi_2(u)\}\oplus\mathcal{H}_3$ estimated via penalized partial likelihood with J_3 as the penalty.

4. Asymptotic Theory

4.1. Linear approximation

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

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_{\mathcal{T}}\eta_i dN_i - \int_{\mathcal{T}}\eta_i Y_i e^{\eta_{0,i}}\Big\} + \frac{1}{2}V(\eta-\eta_0) + \frac{\lambda}{2}J(\eta), \tag{4.1}$$

where $\eta_{0,i}(t) = \eta_0(t,U_i)$. η_1 is linear in dN_i. 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 (4.1) and solving for $\eta_{\nu,1}$, one obtains $\eta_{\nu,1} = (\beta_\nu + \eta_{\nu,0})/(1+\lambda\rho_\nu)$, where $\beta_\nu = (1/n)\sum_{i=1}^{n}\int_{\mathcal{T}}\phi_{\nu,i}dM_i$ and $\phi_{\nu,i}(t) = \phi_\nu(t,U_i)$. By (2.3), (2.4), and noting that $\int_{\mathcal{U}}\int_{\mathcal{T}}\phi_\nu^2 e^{\eta_0}\tilde{S}m = V(\phi_\nu) = 1$, $E\beta_\nu = 0$ and $E\beta_\nu^2 = n^{-1}$. It then follows that

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

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

as $n \to \infty$ and $\lambda \to 0$ (see Gu and Qiu (1993), Theorem 5.1 and also Silverman (1982), §6).

Theorem 4.1. *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* $\lambda J(\eta_1 - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

4.2. Approximation error

Let $\hat\eta$ be a minimizer of (2.6). Let $L(\eta) = -(1/n)\sum_{i=1}^n \{\int_{\mathcal{T}} \eta_i dN_i - \int_{\mathcal{T}} Y_i e^{\eta_i}\}$ and $B_{\eta,h}(\alpha) = L(\eta + \alpha h) + (\lambda/2)J(\eta + \alpha h)$. It can be shown that

$$0 = \dot B_{\hat\eta,\hat\eta-\eta_1}(0) = -\frac{1}{n}\sum_{i=1}^n \Big\{ \int_{\mathcal{T}} (\hat\eta - \eta_1)_i dN_i - \int_{\mathcal{T}} (\hat\eta - \eta_1)_i Y_i e^{\hat\eta_i} \Big\} + \lambda J(\hat\eta, \hat\eta - \eta_1). \tag{4.2}$$

Similarly, define $L_1(\eta) = (1/n)\sum_{i=1}^n \{\int_{\mathcal{T}} \eta_i dN_i - \int \eta_i Y_i e^{\eta_{0,i}}\} + (1/2)V(\eta - \eta_0)$ and $C_{\eta,h}(\alpha) = L_1(\eta + \alpha h) + (\lambda/2)J(\eta + \alpha h)$. It follows that

$$\begin{aligned} 0 = \dot C_{\eta_1,\hat\eta-\eta_1}(0) = -\frac{1}{n}\sum_{i=1}^n \Big\{ \int_{\mathcal{T}} (\hat\eta - \eta_1)_i dN_i - \int_{\mathcal{T}} (\hat\eta - \eta_1)_i Y_i e^{\eta_{0,i}} \Big\} \\ + V(\eta_1 - \eta_0, \hat\eta - \eta_1) + \lambda J(\eta_1, \hat\eta - \eta_1). \end{aligned} \tag{4.3}$$

In equating (4.2) and (4.3), some algebra yields

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} (\hat\eta - \eta_1)_i (e^{\hat\eta} - e^{\eta_1})_i Y_i + \lambda J(\hat\eta - \eta_1) \\ &= V(\eta_1 - \eta_0, \hat\eta - \eta_1) - \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} (\hat\eta - \eta_1)_i (e^{\eta_1} - e^{\eta_0})_i Y_i. \end{aligned} \tag{4.4}$$

One needs the following technical assumptions in order 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 $c_1 e^{\eta_0} \le e^{\eta} \le c_2 e^{\eta_0}$ uniformly on $\{(t,u) : \tilde S(t,u) > 0\}$.

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

Assumption A.4. $\exists c_3 < \infty$ such that $\int_{\mathcal{U}} \int_{\mathcal{T}} \phi_\nu^2 \phi_\mu^2 e^{k\eta_0} \tilde S m \le c_3$, $k = 1, 2$, $\forall \nu, \mu$.

When η_0 is bounded, A.4 essentially requires a uniform bound on the fourth moments of ϕ_ν. By A.3,

$$c_1 \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} (\hat\eta - \eta_1)_i^2 e^{\eta_{0,i}} Y_i \le \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} (\hat\eta - \eta_1)_i (e^{\hat\eta} - e^{\eta_1})_i Y_i. \tag{4.5}$$

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

$$
\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat\eta-\eta_1)_i^2 e^{\eta_{0,i}}Y_i - V(\hat\eta-\eta_1)\Big|\\
&=\Big|\sum_{\nu}\sum_{\mu}(\hat\eta_\nu-\eta_{\nu,1})(\hat\eta_\mu-\eta_{\mu,1})\{\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\phi_{\nu,i}\phi_{\mu,i}e^{\eta_{0,i}}Y_i-\int_{\mathcal{U}}\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\}\Big|\\
&\le\Big\{\sum_{\nu}\sum_{\mu}(1+\lambda\rho_\nu)(1+\lambda\rho_\mu)(\hat\eta_\nu-\eta_{\nu,1})^2(\hat\eta_\mu-\eta_{\mu,1})^2\Big\}^{1/2}\\
&\qquad\Big\{\sum_{\nu}\sum_{\mu}(1+\lambda\rho_\nu)^{-1}(1+\lambda\rho_\mu)^{-1}\{\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\phi_{\nu,i}\phi_{\mu,i}e^{\eta_{0,i}}Y_i\\
&\qquad\qquad-\int_{\mathcal{U}}\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\}^2\Big\}^{1/2}\\
&=(V+\lambda J)(\hat\eta-\eta_1)O_p(n^{-1/2}\lambda^{-1/r}),
\end{aligned}
\tag{4.6}
$$

where Cauchy-Schwartz inequality,

$$
E\Big\{\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\phi_{\nu,i}\phi_{\mu,i}e^{\eta_{0,i}}Y_i-\int_{\mathcal{U}}\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\Big\}^2=O(n^{-1})
\tag{4.7}
$$

via A.4, and $\sum_\nu(1+\lambda\rho_\nu)^{-1}=O(\lambda^{-1/r})$ (Gu and Qiu (1993), Lemma 5.2) are used. To verify (4.7), note that

$$
\begin{aligned}
&E\Big\{\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}Y-\int_{\mathcal{U}}\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\Big\}^2\\
&=E\Big\{\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}(Y-\tilde{S})\Big\}^2+E\Big\{\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}\tilde{S}-\int_{\mathcal{U}}\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\Big\}^2\\
&\le E\Big\{\int_{\mathcal{T}}|\phi_\nu\phi_\mu|e^{\eta_0}\tilde{S}^{1/2}\Big\}\Big\{\int_{\mathcal{T}}|\phi_\nu\phi_\mu|e^{\eta_0}\tilde{S}^{-1/2}E[(Y-\tilde{S})^2|U]\Big\}+E\Big\{\int_{\mathcal{T}}\phi_\nu\phi_\mu e^{\eta_0}\tilde{S}\Big\}^2\\
&\le E\Big\{\int_{\mathcal{T}}|\phi_\nu\phi_\mu|e^{\eta_0}\tilde{S}^{1/2}\Big\}^2+\int_{\mathcal{U}}\int_{\mathcal{T}}\phi_\nu^2\phi_\mu^2e^{2\eta_0}\tilde{S}^2m\\
&\le 2c_3.
\end{aligned}
\tag{4.8}
$$

Similar to (4.5) and (4.6),

$$
\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat\eta-\eta_1)_i(e^{\eta_1}-e^{\eta_0})_iY_i=c\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\eta_1-\eta_0)_i(\hat\eta-\eta_1)_ie^{\eta_{0,i}}Y_i,
\tag{4.9}
$$

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

$$
\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\eta_1-\eta_0)_i(\hat\eta-\eta_1)_ie^{\eta_{0,i}}Y_i-V(\eta_1-\eta_0,\hat\eta-\eta_1)\Big|\\
&=\ (V+\lambda J)^{1/2}(\eta_1-\eta_0)(V+\lambda J)^{1/2}(\hat\eta-\eta_1)O_p(n^{-1/2}\lambda^{-1/r}).
\end{aligned}
\tag{4.10}
$$

Combining (4.4) – (4.10) and letting $n\lambda^{2/r} \to \infty$,

$$\begin{aligned}&(c_1V + \lambda J)(\hat{\eta} - \eta_1)(1 + o_p(1))\\ &\leq |c-1|V(\eta_1-\eta_0, \hat{\eta}-\eta_1) + (V+\lambda J)^{1/2}(\eta_1-\eta_0)(V+\lambda J)^{1/2}(\hat{\eta}-\eta_1)o_p(1). \quad (4.11)\end{aligned}$$

Theorem 4.2. *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 (4.11), Cauchy-Schwartz inequality and Theorem 4.1.

Theorem 4.3. *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* $\mathrm{SKL}(\eta_0, \hat{\eta}) = O_p(n^{-1}\lambda^{-1/r} +\lambda)$.

Theorem 4.3 is a direct consequence of Theorems 4.1, 4.2, and Assumption A.3.

4.3. Semiparametric adaptive estimate

Let $\hat{\eta}_n$ be a minimizer of (2.6) in $\mathcal{H}_n = J_\perp \oplus \{R_J((X_i, U_i), \cdot), \delta_i = 1\}$. I now show that $\hat{\eta}_n$ shares the same asymptotic convergence rates as $\hat{\eta}$.

Let $h \in \mathcal{H}\ominus\mathcal{H}_n \subset \mathcal{H}\ominus J_\perp$. It follows that $\delta_i h(X_i, U_i) = \delta_i J(R_J((X_i, U_i), \cdot), h) = 0$. So $\sum_{i=1}^n \int h_i^2 dN_i = \sum_{i=1}^n \delta_i h^2(X_i, U_i) = 0$ where $h_i(t) = h(t, U_i)$. By (2.3) – (2.5), A.4, and (4.8), $E\{\int_{\mathcal{T}} \phi_\nu\phi_\mu dN - \int_{\mathcal{U}}\int_{\mathcal{T}} \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\} = 0$ and

$$\begin{aligned}&E\Big\{\int_{\mathcal{T}} \phi_\nu\phi_\mu dN - \int_{\mathcal{U}}\int_{\mathcal{T}} \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\Big\}^2\\ &= \int_{\mathcal{U}}\int_{\mathcal{T}} \phi_\nu^2\phi_\mu^2 e^{\eta_0}\tilde{S}m + E\Big\{\int_{\mathcal{T}} \phi_\nu\phi_\mu e^{\eta_0} Y - \int_{\mathcal{U}}\int_{\mathcal{T}} \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\Big\}^2 \leq 3c_3. \quad (4.12)\end{aligned}$$

Lemma 4.1. *Under* A.1, A.2, A.4, *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 (4.6),

$$\begin{aligned}V(h) &= \Big|\sum_\nu\sum_\mu h_\nu h_\mu\Big\{\frac{1}{n}\sum_{i=1}^n\int_{\mathcal{T}} \phi_{\nu,i}\phi_{\mu,i}dN_i - \int_{\mathcal{U}}\int_{\mathcal{T}} \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\Big\}\Big|\\ &= (V + \lambda J)(h)O_p(n^{-1/2}\lambda^{-1/r}),\end{aligned}$$

where (4.12) is used to bound $E\{\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} \phi_{\nu,i}\phi_{\mu,i}dN_i - \int_{\mathcal{U}}\int_{\mathcal{T}} \phi_\nu\phi_\mu e^{\eta_0}\tilde{S}m\}^2$.

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

$$\lambda J(\hat{\eta} - \eta_n) = \frac{1}{n}\sum_{i=1}^n\int_{\mathcal{T}}(\hat{\eta} - \eta_n)_i dM_i - \frac{1}{n}\sum_{i=1}^n\int_{\mathcal{T}}(\hat{\eta} - \eta_n)_i(e^{\hat{\eta}} - e^{\eta_0})_i Y_i. \quad (4.13)$$

Applying the technique used in (4.6),

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\eta_n)_i dM_i\Big| = \Big|\sum_{\nu}(\hat{\eta}_\nu-\eta_{\nu,n})\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\phi_{\nu,i}dM_i\Big|$$
$$= (V+\lambda J)^{1/2}(\hat{\eta}-\eta_n)O_p(n^{-1/2}\lambda^{-1/2r}). \quad (4.14)$$

Similar to (4.9) and (4.10), letting $n\lambda^{2/r}\to\infty$ and using A.3 and Lemma 4.1,

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\eta_n)_i(e^{\hat{\eta}}-e^{\eta_0})_iY_i\Big| = (\lambda J)^{1/2}(\hat{\eta}-\eta_n)(V+\lambda J)^{1/2}(\hat{\eta}-\eta_0)o_p(1) \quad (4.15)$$

Theorem 4.4. *Under* A.1 – A.4, *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 4.4 follows from (4.13) – (4.15) and Theorem 4.3.

I 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

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}_n-\eta_n)_i(e^{\hat{\eta}_n}-e^{\eta_n})_iY_i+\lambda J(\hat{\eta}_n-\eta_n)+\lambda J(\hat{\eta}-\eta_n)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\eta_n)_i dM_i+\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}_n-\eta_n)_i(e^{\hat{\eta}}-e^{\eta_n})_iY_i$$
$$+\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\eta_n)_i(e^{\eta_0}-e^{\hat{\eta}})_iY_i. \quad (4.16)$$

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

$$c_1V(\hat{\eta}_n-\eta_n)+(V+\lambda J)(\hat{\eta}_n-\eta_n)o_p(1)\le\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}_n-\eta_n)_i(e^{\hat{\eta}_n}-e^{\eta_n})_iY_i, \quad (4.17)$$

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}_n-\eta_n)_i(e^{\hat{\eta}}-e^{\eta_n})_iY_i\Big| = (V+\lambda J)^{1/2}(\hat{\eta}_n-\eta_n)(\lambda J)^{1/2}(\hat{\eta}-\eta_n)o_p(1), \quad (4.18)$$

and

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\eta_n)_i(e^{\eta_0}-e^{\hat{\eta}})_iY_i\Big| = (V+\lambda J)^{1/2}(\hat{\eta}-\eta_0)(\lambda J)^{1/2}(\hat{\eta}-\eta_n)o_p(1). \quad (4.19)$$

Combining (4.16) – (4.19) and (4.14), and substituting in the results of Theorems 4.3 and 4.4,

$$(c_1V+\lambda J)(\hat{\eta}_n-\eta_n)(1+o_p(1))$$
$$\le(V+\lambda J)^{1/2}(\hat{\eta}_n-\eta_n)o_p(n^{-1/2}\lambda^{-1/2r}+\lambda^{1/2})+O_p(n^{-1}\lambda^{-1/r}+\lambda). \quad (4.20)$$

This proves the following theorem.

Theorem 4.5. *Under* A.1 – A.4, *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 4.3, 4.4, 4.5, and Assumption A.3.

Theorem 4.6. *Under* A.1 – A.4, *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* $\mathrm{SKL}(\eta_0, \hat{\eta}_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

5. More on Examples

A.3 and A.4 are typical of technical regularity conditions, which in general are very difficult to verify from more primitive conditions. Nevertheless, they appear to be highly plausible as I shall discuss briefly. If η_0 is bounded, then functions in a ball around η_0 of constant radius have uniform bounds, so A.3 may fail to hold only when η_1, $\hat{\eta}$, or $\hat{\eta}_n$ systematically move away from η_0 as $n \to \infty$. In general, ϕ_ν are not available in explicit forms and their fourth moments not computable. Nevertheless, ϕ_ν represent more wiggliness or higher frequencies as $\nu \to \infty$, and there is no special reason that their magnitudes should grow indefinitely, so A.4 appears to be mild. In an overly simplified suggestive special case of Example 1 where $e^{\eta_0}\tilde{S} \propto 1$ and $\mathcal{H} = \{\eta : \int_0^1 \ddot{\eta}^2 < \infty, \eta \text{ periodic}\}$, ϕ_ν are the familiar sinusoidal functions and A.4 follows trivially.

I assume A.3 and A.4 in the remainder of this section and discuss further aspects of the examples of Section 3.

Example 1. *Singleton* $\mathcal{U}$ (continued). It is clear from the theory of Section 4 that the convergence rates of $\hat{\eta}$ and $\hat{\eta}_n$ are $O_p(n^{-1}\lambda^{-1/4} + \lambda)$. The best rates $O_p(n^{-4/5})$ are attained when $\lambda = O(n^{-4/5})$, which satisfies $n\lambda^{2/4} = O(n^{3/5}) \to \infty$. To calculate $\hat{\eta}_n$, one needs to identify R_J. If the norm in $J_\perp$ is taken as $(\int_0^1 \eta)^2 + (\int_0^1 \dot{\eta})^2$, then the reproducing kernel of $\mathcal{H} \ominus J_\perp = \{\eta : \int_0^1 \ddot{\eta}^2 < \infty, \int_0^1 \eta = \int_0^1 \dot{\eta} = 0\}$ with a square norm $\int_0^1 \ddot{\eta}^2$ is

$$R_J(t,s) = K(t,s) \overset{\text{def}}{=} k_2(t)k_2(s) - k_4(|t-s|),$$

where $k_\nu = B_\nu/\nu!$ and B_ν is the νth Bernoulli polynomial (see, e.g., Craven and Wahba (1979)).

Example 2. *Doubleton* $\mathcal{U}$ (continued). Similar to Example 1, the convergence rates are $O_p(n^{-1}\lambda^{-1/4} + \lambda)$. I shall now identify a R_J for calculating $\hat{\eta}_n$. With a square norm $J(\eta) = \theta_1^{-1}J_1 + \theta_2^{-1}J_2$ in $\mathcal{H}_1 \oplus \mathcal{H}_2$, the reproducing kernel can be shown to be

$$\begin{aligned} R_J((t,u),(s,v)) &= \theta_1 R_1((t,u),(s,v)) + \theta_2 R_2((t,u),(s,v)) \\ &= \theta_1 K(t,s)\omega_1(u)\omega_1(v) + \theta_2 K(t,s)\omega_2(u)\omega_2(v), \end{aligned}$$

where $R_1((t,u),\cdot)$ and $R_2((t,u),\cdot)$ span $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively.

Example 3. $\mathcal{U} = [0,1]$ (continued). When $\mathcal{H}_5$ is part of the model space, the convergence rates are seen to be $O_p(n^{-1}\lambda^{\epsilon-1/4} + \lambda)$, $\forall\epsilon > 0$, and the best attainable rates are $O_p(n^{\epsilon-4/5})$, $\forall\epsilon > 0$. When $\mathcal{H}_5$ is not part of the model space, as is the case with the proportional hazard model and the model of Zucker and Karr (1990), the rates are simply $O_p(n^{-1}\lambda^{-1/4} + \lambda)$.

$J_\beta(\eta)$ are square norms in $\mathcal{H}_\beta$, $\beta = 1, \ldots, 5$, respectively, and the associated reproducing kernels are $R_1((t,u),(s,v)) = K(t,s)\psi_1(u)\psi_1(v)$, $R_2((t,u),(s,v)) = \psi_1(t)\psi_1(s)K(u,v)$, $R_3((t,u),(s,v)) = K(t,s)\psi_2(u)\psi_2(v)$, $R_4((t,u),(s,v)) = \psi_2(t)\psi_2(s)K(u,v)$, and $R_5((t,u),(s,v)) = K(t,s)K(u,v)$. Taking $J(\eta) = \sum_{\beta=1}^{5}\beta_\beta^{-1}J_\beta(\eta)$ as the square norm in $\oplus_{\beta=1}^{5}\mathcal{H}_\beta$, the reproducing kernel is $R_J = \sum_{\beta=1}^{5}\theta_\beta R_\beta$. Setting $\theta_\beta = 0$ eliminates $\mathcal{H}_\beta$ from the model space.

The construction of reproducing kernels in Examples 2 and 3 are instances of the construction of tensor-product reproducing kernel Hilbert spaces (see, e.g., Aronszajn (1950), Wahba (1990), and Gu and Wahba (1991)). The extra smoothing parameters θ_β control the relative loads of the penalty $\lambda J(\eta)$ on the roughness of individual terms.

The calculation of $\hat{\eta}_n$ in Example 1 with an automatic λ was developed and illustrated in Gu (1994) and portable code is available. The algorithm is generic and applies also to $\hat{\eta}_n$ in Examples 2 and 3 if subjective choices of θ_β can be made. For the calculation of $\hat{\eta}_n$ in Examples 2 and 3 with automatic objective multiple smoothing parameters (λ and θ_β), an algorithm with data examples has been explored in Gu (1995).

6. Remarks

Compared to previous analyses of closely related problems, the theory developed in this article is more general yet simpler. The simplicity comes from the choice of the natural loss functions $\mathrm{SKL}(\hat{\eta},\eta_0)$ and $V(\hat{\eta}-\eta_0)$ and the indirect analysis of the indirectly defined estimates. Note that no attempt is made to explicitly express $\hat{\eta}$ or $\hat{\eta}_n$, which may not even be unique. Only the minimizing properties are employed in the analysis.

In the absence of a covariate, the convergence rates of $\hat{\eta}$ in $\mathrm{SKL}(\hat{\eta},\eta_0)$ and $V(\hat{\eta}-\eta_0)$ may be obtained from the analyses of Antoniadis (1989) and Cox and O'Sullivan (1990), but those of an adaptive $\hat{\eta}_n$ appear to be new. In the presence of a covariate, the estimates proposed did not seem to exist in the literature, and a computable $\hat{\eta}_n$ makes the first step towards the practicability of the methodology.

The examples presented are among the simplest possible hazard models obtainable via the methodology in that the covariate domains are the simplest possible. When the covariate u consists of several qualitatively different components, possibly a mixture of categorical and continuous ones, the theory remains intact whereas further structures may be introduced on the domain $\mathcal{U}$ via the tensor-product spline technique.

Acknowledgement

This research was supported by National Science Foundation grants.

References

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

Antoniadis, A. (1989). A penalty method for nonparametric estimation of the intensity function of a counting process. *Ann. Inst. Statist. Math.* **41**, 781-807.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337-404.

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.

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

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377-403.

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

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

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255-277.

Gu, C. (1994). Penalized likelihood hazard estimation: Algorithm and examples. In *Statistical Decision Theory and Related Topics V*, (Edited by S. S. Gupta and J. Berger), 61-72.

Gu, C. (1995). Structural multivariate function estimation: Some automatic density and hazard estimates. Technical Report 93-28 (Rev.), Purdue University, Dept. of Statistics.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Ann. Statist.* **21**, 217-234.

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

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

O'Sullivan, F. (1988b). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput.* **9**, 531-542.

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

Utreras, F. D. (1981). Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Statist. Comput.* **2**, 349-362.

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.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Ann. Statist.* **18**, 329-353.

Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.

(Received August 1994; accepted February 1996)